7-23-2015

An Evaluation of the Alignment Method for Detecting Measurement Non-invariance in Noncognitive Scales

Jessica K. Flake
University of Connecticut - Storrs, kayflake@gmail.com

bAn Evaluation of the Alignment Method for Detecting

Measurement Noninvariance in Noncognitive Scales

Jessica Kay Flake, PhD

University of Connecticut, 2015

In recent years a new methodology, the alignment method (Asparouhov & Muthén, 2014), has surfaced for estimating measurement models and detecting measurement noninvariance (i.e., DIF) across many groups. The purpose of the current study was to investigate the alignment method for use with non-cognitive scales across groups of students from different educational contexts (e.g., schools or programs). Asparouhov and Muthén (2014) have investigated the method with continuous and binary item scales, thus I extended previous research by using simulation techniques to evaluate the method with polytomous items, which are often used to measure noncognitive constructs. I also evaluated the new tests of noninvariance produced by the alignment method to a greater extent than has been seen in previous research. Results indicate that the alignment method adequately recovers parameter estimates under small and moderate amounts of noninvariance, with issues only arising in the more extreme conditions. The tests of noninvariance were found to be too conservative for most items, with a near zero Type I error rate. The testing procedure showed appropriate power in polytomous items that were less skewed in the mean structure, which suggests that the psychometric properties of individual items have a large effect on the performance of the procedure.

An Evaluation of the Alignment Method for Detecting Measurement Noninvariance in Noncognitive Scales

Jessica Kay Flake

B.S., Northern Kentucky University, **2010**

M.A., James Madison University, **2012**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

An Evaluation of the Alignment Method for Detecting
Measurement Noninvariance in Noncognitive Scales

Presented by

Jessica Kay Flake, B.S., M.A.

Major Advisor _____
D. Betsy McCoach

Associate Advisor _____
H. Jane Rogers

Associate Advisor _____
Noel Card

University of Connecticut
2015

Acknowledgements

I want to begin by acknowledging Marvin Counts, the counselor at the adult high school in Covington, KY. When I walked into his office, a 17 year old who had just left home, he gave me some hard advice. He told me that it would take me over a year to graduate high school there and that I should get my GED and apply for college. I took his advice, and it allowed me to start college the next year. Along the way I have met so many more people like Marvin, who nudged me in the right direction and encouraged me to continue my education. I want to acknowledge Peg Adams, James Thomas, Josh Greunke, Ross Markle, Kenn Barron, and Chris Hulleman, all of whom mentored me during my bachelor's and master's program and supported me in pursuing my passions.

I also want to acknowledge my friends and family who have helped me through this process. First I want to acknowledge Kay Ridgway, Gail Ohr, Delores Nunnelley, Dawnielle Cunningham, and Cathy Hehman for acting as mothers to me over the years. I also want to thank Eric Hehman, who has gone above and beyond the role of significant other in helping me to develop as a researcher and academic. From writing advice, to watching my practice presentations, to invaluable coauthor, he has inspired me to be a better researcher and motivated me to persist. I also want to thank Matthew Swain, Kelli Samonte, and Erin Strauts, dear friends who helped me format, proof read, and program to get this dissertation finished.

I also want to thank my committee members, Jane Rogers and Noel Card for their guidance and feedback. Finally, I want to acknowledge my advisor Betsy McCoach, who sacrificed inordinate amounts of time to guide me through this process. Sometimes it was her unwavering belief that I could complete this project that motivated me to continue, in the face of my own doubts. I could not have finished this without her.

Table of Contents

Chapter 1: Introduction

In America, educational assessment and measurement are in the midst of an historical change. States are shifting away from traditional multiple-choice assessments and incorporating the measurement of new constructs. These movements for more authentic and broader assessments are necessitating the development of new quantitative methodology. The evaluation of one such methodology is the focus of this study.

A Shift Away from Zeros and Ones

As of early 2015, forty-three states and the District of Columbia had adopted the Common Core State Standards (CCSSI, 2015). This is the first move in American history toward a national set of standards, and consequently a national set of curricula and assessments. Two large consortia have evolved to develop the corresponding assessments: Smarter Balanced and the Partnership for Assessment of Readiness for College and Careers (PARCC). Both of these consortia boast to deemphasize multiple choice items that are scored right (a one) or wrong (a zero). The Smarter Balanced assessments have five item types/tasks: selected response, constructed response, extended response, technology-enhanced, and even performance tasks (Measured Progress & ETS, 2012), all of which will allow for partial credit. The PARCC consortium states explicitly that their items, "*allow students to demonstrate what they know, rather than what they don't know*—where items allow for partial credit" (PARCC, 2015).

Relatedly, there is a drastic increase in measuring student skills beyond academic subjects. Though educators and measurement experts have long understood and spoken to the importance of student development beyond academic subject matter (e.g., social-emotional learning or non-cognitive skills), the past few years have brought the idea into the spotlight. For example, writings such as Tough's (2012) *How Children Succeed: Grit, Curiosity, and the*

Hidden Power of Character have made national news, alongside heated debates about the absence of Common Core standards that address learning outside of academic subjects. In response, some states, such as Illinois and California, have taken an interest in formally teaching and assessing non-cognitive skills alongside the Common Core curriculum. Further, pre-kindergarten educators have taken an official stance on assessing the "whole child" with national standards and assessments that incorporate attitudes and skills beyond numeracy and literacy (National Association for the Education of Young Children, 2009).

Broadly defined, non-cognitive skills are skills or attributes that are not typically measured by cognitive, standardized tests. Though the use of this term is debated in the field (Duckworth & Yeager, 2015) it generally refers to, but is not limited to: personal qualities, attitudes, personality, psycho-social skills, social-emotional learning, emotional intelligence, and motivation. These constructs have also been discussed as 21st century skills (Baker, 2013) which acknowledges their importance as we move forward into a more global and digitized age. Support for teaching and measuring such constructs has been garnered through consistent research demonstrating their importance. There is a wealth of research showing that non-cognitive skills play a crucial role in student performance and persistence, as well as numerous meta-analytic studies (e.g., Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Robbins et al., 2004).

Both types of assessments, partial credit items for achievement tests and noncognitive items, yield ordered, categorical (i.e., polytomous) data. In the case of achievement tests, the shift toward partial credit items means that students will no longer just have zeros or ones as their responses, but a range of potential points. In measuring non-cognitive skills, the items typically utilize a Likert response scale, where the student reads a statement and indicates how

much they agree with each statement. These scales generally range from having four to seven discrete response options. Some research supports that ordered categorical responses with many response options (seven or more) can be treated as continuous in statistical models (Dolan, 1994). However, measures with five or fewer response options are extremely common in educational measurement. For example, noncognitive measures often have a 4 or 5-point "Strongly Agree" to "Strongly Disagree" response scale, as well as partial credit items that allow the student to earn between three and five points. Thus, ordered categorical data analysis is required for these types of assessments.

Methodological Challenges

These changes in educational measurement and practice introduce the capacity to compare different schools across the states, or different states on their achievement and noncognitive skills. But first, we need to consider if it is even reasonable to assume that these measures of achievement and noncognitive skills are comparable across these different educational contexts. This issue is particularly salient to the member states of testing consortia and large testing programs, such as the ACT Engage and the ETS Mission Skills Assessment. These assessments are being administered nationally and the various grade schools, high schools, and colleges are likely to want to know where they rank in promoting these skills. However, making such comparisons presents unique methodological challenges.

The multiple group factor model can be used to compare groups on their factor means (e.g., level of noncognitive skills); however for those comparisons to be valid one must assume measurement invariance. Measurement invariance (i.e., measurement equivalence and factor invariance) describes a characteristic in which the measurement properties of a scale do not change, or are invariant, across contexts or groups of people. Measurement invariance is

implicitly assumed whenever measurement models are used to answer questions about differences in a construct across time or between groups. For example, if studying the differences in a construct between males and females, it is necessary that the measurement of the construct is the same across the genders, such that group differences are not conflated with differences in measurement. Measurement invariance is equally important when scales are used to measure change across time; if the scale properties are also changing across time it is difficult to tease apart the different sources of the change.

Conceptually, measurement invariance for a construct holds when the nature or conceptualization of the construct is the same across different people or different situations. When applying this concept to a non-cognitive construct, like student motivation, we would assume that the items on a motivation scale are conceptualized in the same manner for different types of students. For example, an engineering student would think of an item on a motivation scale in the same way as an art student. This property of measurement empirically manifests in the estimates derived from quantitative analyses.

There are many statistical approaches for testing measurement invariance, but they all share the necessity of an investigation of numerous parts of the measurement model. This is typically a multistep process and throughout one can encounter a number of difficulties. For example, researchers typically conduct tests on each item of a scale, one by one, to discover where in the measurement model noninvariance exists. As such, testing for measurement invariance becomes more onerous as the number of scale items increases. This process becomes even more complex with polytomous data, such as survey data from Likert scales, because one needs to confirm that the use of each option on the response scale is invariant across groups. If

the researcher has many items and many response options this creates hundreds of statistical tests.

One method commonly used to simplify this process is to treat the ordered response categories as a continuum. When this is done, researchers test that the mean of the continuum is the same across groups, instead of investigating each response category individually. Though this reduces the number of statistical tests, treating ordinal data with less than five categories as a continuous scale is generally discouraged in the field. This type of model misspecification can lead to biased results (Dolan, 1994). Further, it can be particularly problematic in the case of testing for measurement invariance, as it is difficult to know if bias in the model is from noninvariance or from the misspecification of treating ordinal data as continuous (Lubke & Muthén, 2009).

Finally, the above challenges are exacerbated and become insurmountable if one is interested in more than two or three groups. The standard procedure of testing each item at a time for invariance would need to be conducted for every possible pair of groups manually. This results in a myriad of statistical tests where errors of inference (e.g., Type I, Type II, over-fitting) are likely. In addition to these tests being tedious and time consuming, the higher probability of making a Type I or Type II error make it unlikely that one would arrive at the true model, uncovering the real pattern of noninvariance. Further, iterative model re-specification is completely data driven, and this process capitalizes on chance, such that findings are unlikely to generalize to other samples (MacCallum, Roznowski, & Necowitz, 1992).

These methodological challenges have limited researchers to conduct invariance tests with a limited number of groups, where membership is defined by person characteristics, such as gender, race, or nationality. But what shapes the conceptualization of a construct and causes

noninvariance? Noninvariance can be caused by a person's experience, and their situation. This foundational, psychological principle is shown in Kurt Lewin's (1931) famous equation, where behavior is a function of the person and the situation:

$$\text{Behavior} = f(\text{Person}, \text{Situation})$$

In the context of measurement invariance, the item response is the behavior, and traditional tests for noninvariance have focused on the person portion of the equation. When considering achievement and noncognitive development of students and student success, educators and policy makers are interested in effective programs, high-impact educational practice, and what interventions work. These questions are not about the person; they are about the situation and necessitate the study of invariance across many groups.

Recent advances in quantitative methodology allow for the consideration of larger numbers of groups, introducing the possibility to investigate noninvariance that arises from the situation. Further, these methods can accommodate ordinal data and simplify the process of detecting noninvariance. The purpose of this dissertation is to investigate a new statistical procedure, the alignment method, for meeting the challenges associated with making comparisons across many groups when data are polytomous. First, the alignment method can accommodate polytomous data, which is common in achievement tests and noncognitive and psychological assessments. Second, to compare groups' scores on such scales one must assume that the scales function in the same way across groups (i.e., are invariant), which is assessed by the alignment method. The focus of the dissertation research is to evaluate the viability of the alignment method for assessing invariance of ordered categorical response scales across many groups and identify under what conditions of partial invariance groups are no longer comparable.

Current simulation studies investigating the alignment method show that it works well with scales that have continuous items (Asparouhov & Muthén, 2014). Asparohouv and Muthén (2014) also illustrate the method's potential for handling large scale assessments using a subscale from the European Social Survey. This illustration shows how the method completely automates the task of estimating measurement models across many groups and identifies items or groups that are noninvariant. This could change the way large scale assessment is conducted by replacing the traditional, more arduous techniques, which would make many group comparisons more accessible and valid. This study will directly examine the method's performance with polytomous data and provide recommendations for its use with a large scale noncognitive assessment.

Research Questions

The overall purpose of the dissertation research was to evaluate the alignment method for use in investigating noninvariance of scales comprised of polytomous items. To facilitate this, I conducted a simulation study with some conditions similar to those of Asparohouv and Muthén (2014). However, I extended the research by incorporating additional conditions and analyses of the results. The broad research questions were:

1. How do the number of groups, percentage and magnitude of noninvariance affect bias and coverage of the item parameter estimates?
2. How do the number of groups, percentage and magnitude of noninvariance affect the accuracy of the alignment's post-hoc noninvariance testing?
3. How do the number of groups, percentage and magnitude of noninvariance affect the alignment's recovery of group specific factor means and variances?

The primary focus of the dissertation is practical: to understand under what circumstances the alignment method should be used. As with any new method, researchers need procedures to detect violations of the assumptions and guidance about how those violations will impact their results. For the alignment method, it is important for researchers to know at what number of noninvariant items they should forfeit comparing group means or how different groups can be before results are biased. I will investigate the bias in estimates and accuracy of statistical tests under a variety of magnitudes of noninvariance to facilitate these types of recommendations.

Chapter 2: Literature Review

Factor Analysis and Polytomous Data

A large majority of constructs in the fields of psychology and education are unobservable or latent. For example, one cannot directly witness a student's level of motivation or math ability. However, one can infer the level of these constructs by asking the student questions (i.e., items) on tests or surveys. Statistical models are used to relate these observed responses to the latent construct of interest. A widely employed statistical model for measurement is factor analysis, which uses observable proxies, or item responses, as indicators of the latent, unobservable factor.

There are two broad types of factor analysis: exploratory and confirmatory. Exploratory factor analysis (EFA) is used when the researcher needs to determine the factor structure of the items. In EFA a rotation algorithm (see Browne, 2001 for an overview) is used to uncover the pattern of factors and the relationships the items have to the factors. The desirable solution in EFA is one where items relate strongly to one factor, but not to any other factors, also called simple structure. EFA is generally used in situations where researchers are unsure of how many factors exist and which items relate to them, whereas confirmatory factor analysis (CFA) is used when the researcher has a priori hypotheses about the factor structure and wants to confirm and examine those hypotheses.

To consider how groups and their measurement models might differ we must start with a confirmatory single factor model. For a single group where the items are continuous this model can be represented as:

$$y_{ij} = \nu_{jl} + \sum_{l=1}^{L} \lambda_{jl}\eta_{il} + \varepsilon_{ij} \qquad (1)$$

Equation 1 displays the factor model as a linear regression of the observed indicators on latent factors, where Y_{ij} is the score for person *i* on item *j*, ν_{jl} is the expected value of the item when the corresponding factor, *l*, is zero (i.e., intercept), and λ_{jl} is the regression slope (i.e., loading) from the regression of Y_j on the factor score η_l. ε_{ij} is the residual score of person *i* on item Y_j. Thus, each item's variance is decomposed into two parts, the variance that is shared with the factor and the residual, or error variance that is unique to the item. The multiple group confirmatory factor model is expressed in the same way, except incorporates a subscript to represent each group:

$$y_{ijg} = \nu_{jg} + \sum_{l=1}^{L} \lambda_{jlg}\eta_{ilg} + \varepsilon_{ijg} \qquad (2)$$

In the multiple group factor model (Equation 2), Y_{ijg} is the observed score for person *i* on item *j*, who is a member of group *g*. In these models, the means of the items and covariances among the items are estimated simultaneously for all groups.

Extending to Polytomous Items

We can elaborate on this multiple group factor model to incorporate items that are ordered categorical. In these models, the factor is still a function of responses to the observed variables, but the observed variables are now partitioned into categories of each item:

$$y^*_{ijg} = \nu_{jg} + \sum_{l=1}^{L} \lambda_{jlg}\eta_{ilg} + \varepsilon_{ijg} \qquad (3)$$

$$y_{ij} = c \quad \text{if} \quad \tau_{c-1} < y^*_{ijg} \leq \tau_c .$$

This model (Equation 3) is similar to the model with continuous items, but *y** is not observed. It is assumed that an unobserved continuum underlies the ordered categorical responses. The continuum of *y** is divided by thresholds, τ, for *C* number of categories. The thresholds are

ordered such that $\tau_0 < \tau_1 <\ldots, \tau_c$, with τ_0 equal to $-\infty$ and τ_c equal to $+\infty$. The threshold is interpreted as the amount of the factor needed to transition into the next category. Items will have one fewer thresholds than the number of response options. To model the relationship between η and y^* there are usually three steps: (a) compute the thresholds, (b) estimate the unobserved correlations among y^* variables, (c) estimate the other parameters using the thresholds and latent correlations using weighted least squares estimation (Finney & DiStefano, 2013). Because the indicators are categorical, the factor model with polytomous items is not linear, as is the model with continuous items. The link function that relates the items to the factors is either a logistic or standard normal distribution function, corresponding to logistic or probit regression, respectively (Muthén & Asparouhov, 2013a).

Since y^* is not observed, the scale of it must be set. There are two options for parameterization: standardized and unstandardized. In Mplus, the software featuring the alignment method, these are called DELTA and THETA, respectively. The DELTA scaling fixes the total variance of y^* to one and the residual variance is not estimated as a parameter in the model, but is calculated as the remainder of $1-\lambda^2\Psi$. This scaling method results in a standardized solution when the factor variance is fixed to 1, such that the thresholds can be interpreted as z-scores and the loadings as the standard deviation change in y^* as the factor increases by 1. The THETA scaling produces an unstandardized solution where the residual variance is fixed to 1 and the total variance is not estimated as a parameter in the model, but computed as the sum of residual variance and common variance, $\lambda^2\Psi$. THETA parameterization produces thresholds and loadings that are in a probit regression metric. The probit metric, short for probability unit, is a normal score, which corresponds to a particular probability in the standard normal cumulative distribution function (CDF). For example, to obtain the probability

a correct response for a binary item with a threshold of -2.89 using the THETA parameterization, one would need to reference a CDF table and for the corresponding probability, which is .0026. As such, higher thresholds indicate a higher probability of transitioning into the next response category when the latent factor is zero, or at its mean.

Relating the Common Factor Model to Item Response Theory

Item response theory (IRT) is a common framework for assessing items and tests in educational measurement. IRT encompasses a set of latent variable models that typically involve items that are binary or polytomous. As with factor analysis, IRT models link items to an underlying latent trait through a mathematical function. The IRT model in which the thresholds and loadings are estimated, also called the 2PL model, is represented in Equation 4.

$$P(Y_j = 1 | L) = \frac{1}{1 + \exp[-a_j(\mathrm{L} - b_j)]'} \quad (4)$$

Here the probability of a 1, or of a correct response on cognitive tests, is modeled as a function of *L*, the latent factor, referred to as θ in most IRT models. The item is linked to the factor with a logistic function, as is often done with the common factor model. The discrimination parameter, akin to the loading in factor analysis is *a* for item *j*. The difficulty parameter, akin to the threshold, is *b* for item *j*. The standard normal cumulative distribution function is also used in IRT models and is referred to as the normal ogive. It is approximately equivalent to the logit model when the logit is multiplied by a constant of 1.7, often represented as *D*:

$$P(Y_j = 1 | L) = \frac{1}{1 + \exp[-Da_j(\mathrm{L} - b_j)]'} \quad (5)$$

These models are empirically indistinguishable from the common factor model with binary items (Millsap, 2011). Lord and Novick (1968) demonstrated the connection between the normal ogive model and the common factor model with binary items:

$$a_j = \lambda_{j,} \quad b_j = \frac{\tau_j}{\lambda_j} \tag{6}$$

When the latent factor is scaled with mean = 0 and variance= 1, the discrimination for item *j* is equal to the loading for item *j* , and the difficulty for item *j* is equal to the threshold for item *j* divided by the loading for item *j*. The model presented in equation 5, which relates polytomous item responses to a latent factor in the common factor model, is parameterized in the same way as Samejima's (1997) graded response model often discussed in the IRT framework (Muthén & Asparouhov, 2013a). Therefore, the discussion of the polytomous factor model and empirical investigation of the polytomous alignment presented in this study are relevant to IRT based research and assessment.

Testing for Measurement Noninvariance

Quantitative investigations of measurement invariance, of which there are many, can take on different forms under different frameworks. In this section I will focus on approaches undertaken in latent variables models (SEM and IRT) because they are the most relevant to the genesis of the alignment method. Tests of measurement noninvariance, termed differential item functioning (DIF) detection in IRT models, take on a similar form under a general latent variable framework (Reise, Widaman, & Pugh, 1993). Though these existing approaches are vastly different from the alignment method, they are relevant to the discussion in that they establish the previous context in which the alignment was developed. As I will discuss, the mechanics of the existing approaches and the complications that occur when using them give rise to the rationale for the alignment method.

Traditional Methods of Invariance Testing

The various approaches to testing measurement invariance share the necessity of an investigation of numerous parts of the measurement model. This is typically a multistep process

where each item or sets of items are tested for equality across two groups. A common approach taken in a latent variable framework is to run a sequence of models that increase in equality constraints across groups while assessing change in model fit. This is done using chi-square difference testing (Kline, 2011; Raykov & Marcoulides, 2006), referred to as likelihood ratio testing in an IRT framework (Thissen, Steinberg, & Gerrard, 1986).

In these tests, the difference in fit statistics, the chi-square statistic or log-likelihood, of the simpler model (where item parameters are constrained to be equal across groups) is compared to a more complex model (where item parameters vary across groups). With continuous items, this difference in chi-square statistics or log-likelihoods is chi-square distributed, with degrees of freedom that represent the change in the number of estimated parameters across the two models. If the difference is significant, then the more complex model, where item parameters are estimated separately for each group, should be favored, as it fits the data significantly better. If the model fits significantly better when item parameters are different across groups, the items are deemed noninvariant or differentially functioning.

The general method of comparing nested models to assess noninvariance is discussed in detail below. Earlier writings by Joreskog (1971), in an SEM framework, and Thissen et al., (1986), in an IRT framework, proposed this general type of method. Since, it has been further discussed in numerous illustrative and comparative writings (Horn & McArdle, 1992; Millsap, 2011; Stark, Chernyshenko, & Drasgow, 2006; Vandenberg & Lance, 2000). Below, I outline the process as described by Millsap (2011).

Consider the multiple group factor model with polytomous items:

$$y^*_{ijg} = \nu_{jg} + \sum_{l=1}^{L} \lambda_{jlg}\eta_{ilg} + \varepsilon_{ijg} \qquad (7)$$

$$y_{ij} = c \qquad if \quad \tau_{c-1} < y^*_{ijg} \leq \tau_c .$$

This model can also be represented as two sets of matrices, corresponding to the mean and covariance structure of the items

$$\begin{aligned} \boldsymbol{\mu}^*_{\mathbf{g}} &= \boldsymbol{\tau}_{\mathbf{g}} + \boldsymbol{\Lambda}_{\mathbf{g}}\boldsymbol{\alpha}_{\mathbf{g}} \\ \boldsymbol{\Sigma}^*_{\mathbf{g}} &= \boldsymbol{\Lambda}_{\mathbf{g}}\boldsymbol{\Psi}_{\mathbf{g}}\boldsymbol{\Lambda}_{\mathbf{g}}' + \boldsymbol{\Theta}_{\mathbf{g}} . \end{aligned} \tag{8}$$

The vector of thresholds, $\boldsymbol{\tau}$, factor loadings, $\boldsymbol{\Lambda}$, and the factor mean α, for group *g*, are used to impose structure on the estimated item means, $\boldsymbol{\mu}^*_{\mathbf{g}}$. Structure is also imposed on the covariance matrix of the item responses, $\boldsymbol{\Sigma}^*_{\mathbf{g}}$, where $\boldsymbol{\Lambda}_{\mathbf{g}}$ is the matrix of factor loadings for group *g*, $\boldsymbol{\Psi}_{\mathbf{g}}$ is the covariance matrix of the factor scores, and $\boldsymbol{\Theta}_{\mathbf{g}}$ is the covariance matrix of the residual scores. Again, this shows that the model decomposes the item responses into what is common amongst the items, the factor loadings and factor scores, and what is unique, the residual. The residual is assumed to be uncorrelated with the factor scores.

The logical first step in this process of testing for measurement invariance is to test if the zero elements in the factor structure are the same across groups. This is known as configural invariance (Horn & McArdle, 1992), because it tests if the basic configuration of the factor structure is the same across groups. This hypothesis is represented in the matrix formulation as

$$\begin{aligned} H_{con} = \boldsymbol{\Sigma}^*_{\mathbf{g}} &= \boldsymbol{\Lambda}_{\mathbf{g}}\boldsymbol{\Psi}_{\mathbf{g}}\boldsymbol{\Lambda}_{\mathbf{g}}' + \boldsymbol{\Theta}_{\mathbf{g}}, \\ \boldsymbol{\mu}^*_{\mathbf{g}} &= \boldsymbol{\Lambda}_{\mathbf{g}}\boldsymbol{\alpha}_{\mathbf{g}} . \end{aligned} \tag{9}$$

This level of invariance constrains the general configuration of the factors to be the same across groups, while the loadings, intercepts/thresholds, and error variances are free to vary. Because the loadings and intercepts/thresholds are different across groups, one cannot compare the factor means of groups under configural invariance. However, if there is support for configural

invariance one would then compare the configural model's fit to a model where the factor loadings are constrained to be equal across groups. This is represented as

$$H_{scal} = \boldsymbol{\Sigma}^*_{\mathbf{g}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}_{\mathbf{g}}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Theta}_{\mathbf{g}}, \quad \boldsymbol{\mu}^*_{\mathbf{g}} = \boldsymbol{\Lambda}\ \boldsymbol{\alpha}_{\mathbf{g}} \tag{10}$$

where the factor loadings no longer vary by the group, *g*. If model fit is not substantially compromised then one concludes pattern (Millsap, 2011) or metric (Horn & McArdle, 1992) invariance. Under this level of invariance it has been recommended that one cannot compare the factor means, but can compare mean adjusted statistics such as regression coefficients and covariances. If metric invariance is tenable, one then moves forward to test the invariance of the intercepts/thresholds (i.e., strong factorial invariance, Meredith, 1993; or scalar invariance, Steenkamp & Baumgartner, 1998). Scalar invariance has historically been discussed as necessary for the comparisons of latent factor means and is represented as

$$H_{scal} = \boldsymbol{\Sigma}^*_{\mathbf{g}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}_{\mathbf{g}}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Theta}_{g}, \quad \boldsymbol{\mu}^*_{\mathbf{g}} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\ \boldsymbol{\alpha}_{\mathbf{g}} \tag{11}$$

where the intercepts/thresholds no longer vary across groups. The final step in the process, strict factorial invariance (Millsap, 2011), would test model fit when loadings, intercepts or thresholds, and error variance are constrained to be equal across groups. This final level of measurement invariance allows researchers to conclude that the instruments are equally reliable, as they have the same proportion of common factor variance to error variance across groups. This model is represented as

$$H_{stri} = \boldsymbol{\Sigma}^*_{\mathbf{g}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}_{\mathbf{g}}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Theta}\ , \quad \boldsymbol{\mu}^*_{\mathbf{g}} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\ \boldsymbol{\alpha}_{\mathbf{g}}\ . \tag{12}$$

There are numerous considerations in evaluating the performance of this type of model comparison approach for identifying measurement noninvariance, particularly at the item level. In the literature a few issues have been the focus of methodological research: the selection of invariant items for model comparison and identification, identifying those invariant items, model fit criteria used to make the comparison, and general factors that influence power, such as sample size, amount of noninvariance, and psychometric properties of the items.

Above I described the general method of comparing models discussed by Millsap (2011) where a fully constrained model is tested for each level of invariance. However, it is common that the metric or scalar invariance model for all items would be rejected, indicating at least one noninvariant item is present. Then, one usually wishes to identify which item or items are noninvariant. This can be done a variety of ways and the best procedure for doing so is somewhat debated in the field. Even the notion of testing all items at once for a single level of invariance has been contested. Zumbo and Koh (2005) found that the full metric or scalar model can fail to be rejected when individual items are noninvariant. This means that one can conclude metric or scalar invariance of all items, when in fact some individual items are noninvariant.

In testing for noninvariance of individual items it is necessary to constrain at least one item to be equal across groups to set the scale of the latent variable and identify the model, with an assumption that the constrained item is truly invariant (Johnson, Meade, & DuVernet, 2009), termed an anchor item in IRT. However, there are differences across SEM and IRT in how many items are constrained (Kankaras, Vermunt, & Moors, 2011; Meade, 2004; Stark et al., 2006). In IRT there are numerous methods of anchoring, but it is common to constrain all items to be equal across groups except for the item that is suspected to have DIF. This method, called the "all-other" anchor method works well if the item under investigation is, in fact, the only DIF

item. However, if numerous items are noninvariant, the method lacks power to detect noninvariance (Stark et al., 2006; Wang & Yeh, 2003). There is evidence to support that only one, truly invariant anchor item is needed to detect noninvariance in other items, with appropriate error rates and power, even when noninvariant items are present (Meade & Lautenschlager, 2004; Wang & Yeh, 2003), though performance can be improved with more anchor items.

Then, however, the conundrum becomes identifying which item or items are truly noninvariant. There have been a number of methods and studies related to selecting invariant anchor items and the ramifications of erroneously selecting noninvariant items. Johnson et al., (2009) found that selecting a noninvariant item as the anchor item increased the likelihood of flagging an item with DIF when it is actually invariant. In considering how to select the truly invariant items numerous methods have been proposed and evaluated (e.g., Cheung & Rensvold, 1999; French & Finch, 2008; Woods, 2008). Though a description of these methods is not directly relevant to the alignment, as the alignment does not require selecting anchor items, it is relevant that they all share the necessity of conducting many statistical tests manually. These methods all require testing models again and again with different iterations of anchor and non-anchor items, to hone in on the truly invariant ones.

In addition to considering how many and which items to select as anchors, researchers also need to consider their method of comparing model fit. In this area of research some conflicting results have been found. Numerous studies have supported the use of the chi-square difference test (i.e., likelihood ratio test) for concluding a significant degrade in model fit (Cohen, Kim, & Wollack, 1996; French & Finch, 2006). Whereas others have found that the chi-square difference test was more prone to Type I errors (Cheung & Rensvold, 2002; Rutkowski

& Svetina, 2014), and suggests using a CFI change of .01 to .02 to assess a meaningful change in fit. Based on the results of a simulation Cheung and Rensvold (2002) also suggest using Gamma hat, and McDonald's Non-centrality Index when determining noninvariance. Some have suggested using numerous measures of fit in conjunction (Vandenberg & Lance, 2000), but French and Finch (2006) found that method dramatically decreases power to detect noninvariant items.

Further, there are numerous factors that increase the likelihood of making an error of inference in these tests, the first of which is sample size. Generally latent variable models are large-sample techniques, using estimation (e.g., maximum likelihood) procedures that rely on asymptotically normal properties. For example, French and Finch (2006) had convergence issues with IRT models with sample sizes of less than 300. Adequate power has been observed in studies with sample sizes of 500 to 1000 (Cohen et al., 1996; Meade & Lautenschlager, 2004; Stark et al., 2006). However, French and Finch noted a challenging result: chi-square difference tests had adequate power to detect noninvariance for overall tests, but when testing individual items power was dramatically reduced, especially if there were many noninvariant items.

Sample size is not the only factor that influences power in tests of noninvariance. Important factors to consider are the number of items per factor, number of factors, and the psychometric properties of the scale. Numerous studies have noted that power to detect noninvariance was greatest for the items with largest loadings (Kaplan, 1989; Meade & Bauer, 2007), as well as when there were more items per factor (French & Finch, 2006; Meade & Bauer, 2007). Further, errors of inference are more likely when the factor model is misspecified, such as treating ordinal data as continuous or ignoring cross-loading in the factor structure (French & Finch, 2011; Lubke & Muthén, 2009) .

The Multiple Indicators Multiple Causes Model

Another common model executed in the latent variable framework is the multiple indicators, multiple causes model (MIMIC, Joreskog & Goldberger, 1975). This model is gaining traction for tests of measurement noninvariance (Finch, 2005; Woods, 2009). Figure 1 represents the MIMIC model. This model is the same as the common factor model discussed in the matrix formulation above in that items are linked to a common factor. The factor is represented as the larger circle, labeled factor, with the item loadings as the path that connects the factor to the observed item responses, which are represented with the squares labeled I1-I4. The error variance discussed above is represented as a circle, depicting the variance in the item that is unexplained by the factor. Recall the multiple group model discussed above where a factor model is estimated for each group. Using gender as an example, a model for males and a model for females would be estimated, and differences in the item parameters would then be tested using a chi-square difference test of model fit. The MIMIC model conceptually tests the same hypothesis, but is formulated differently than the multiple group model.

In a MIMIC model, group membership is introduced as a predictor of item and factor variance. If group membership is significantly related to the latent factor, this means that the group significantly different in their factor means. This is often discussed as impact in the IRT literature. If group membership is also related to the item, this means that after controlling for factor means, there is a difference in the predicted mean item response. This is another way of describing measurement noninvariance as it conveys that the predicted mean of the item (i.e., intercept) is different across groups. In the figure, if the item intercept for item four was significantly different across gender, the item would be flagged as noninvariant, or differentially functioning. In considering if the loading of item four is noninvariant, one would then test if

there is a significant interaction between gender and the latent variable. This tests the hypothesis that the slopes across genders are significantly different, which means that after controlling for the factor, the item relates differently for males than for females.

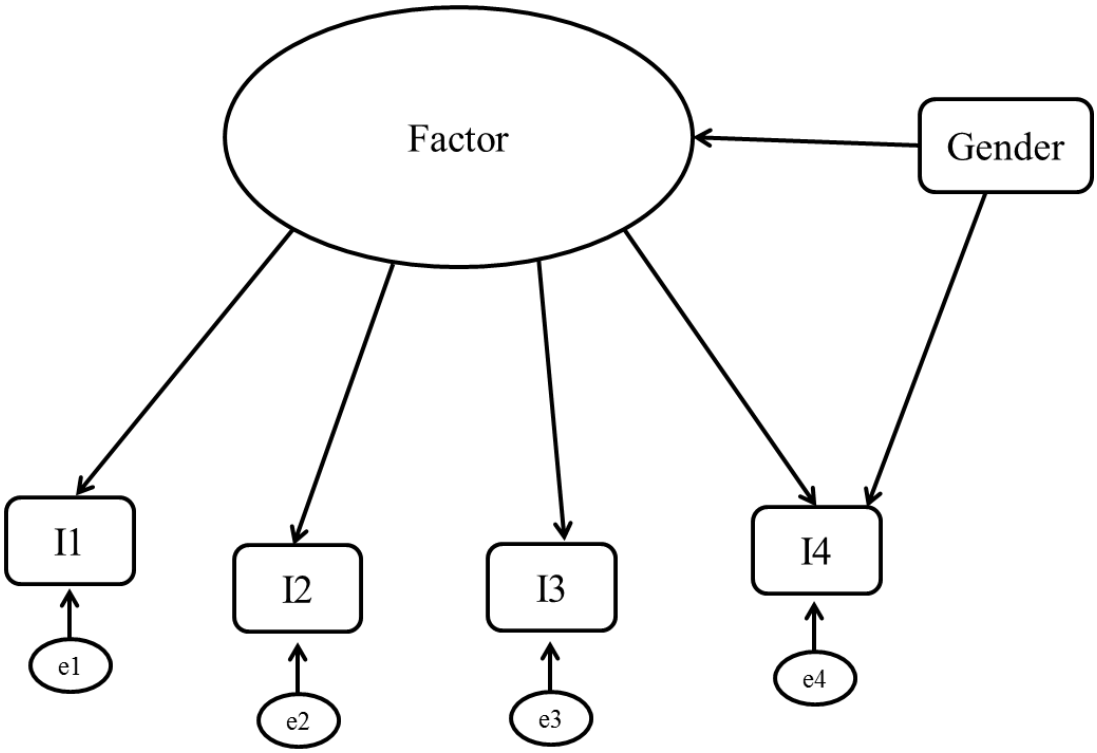

Figure 1. Path Diagram of MIMIC model

Though relatively recent, there have been a number of studies on the efficacy of the MIMIC model to detect noninvariance, particularly in comparison to the chi-square difference testing procedures used in a multiple group model. Numerous studies suggest that the MIMIC model is less prone to Type I errors when detecting noninvariance in large samples than the multiple group approach (Finch, 2005; Woods, 2009). However, the method still suffers from some of the same complications as the multiple group method, such as the need to identify anchor items. There is evidence to suggest the one truly invariant anchor item works best with the MIMIC model (Wang & Yeh, 2003), which has also been noted in the multiple group case (Wang & Yeh, 2003). Further, testing for noninvariance in the loadings can be difficult, as it

introduces the necessity of modeling the interaction between the observed group membership and the latent variable.

Observed Score Methods

There are many other measurement noninvariance testing procedures not described in detail here. Millsap (2011) describes two broad types of procedures: latent variable models and observed score methods. I have focused on latent variable methods, as they are more methodologically connected to the alignment. However, there are a host of noninvariance testing procedures that do not incorporate the estimation of an underlying latent variable. In such methods the observed total test score is used as a proxy. Since latent variable models require large samples, these observed score methodologies are often necessary. Common observed score methods include the Mantel-Haenszel (Holland & Thayer, 1986; Mantel & Haenszel, 1959), the standardization approach (Dorans & Kulick, 1986), and logistic regression (Swaminathan & Rogers, 1990).

Complications with Latent Variable Models

As is evidenced from the simulation work conducted on these methods the researcher can encounter a number of challenges and complications. Many of these challenges revolve around the multiple tests needed to identify which items are invariant and how to proceed if some noninvariance is discovered.

Multiple testing. As mentioned before, to specify any level of invariance, one must work up to that level by testing all applicable parameters at once, then, if that model is rejected via the evaluation of change in goodness of fit indices, go through a series of parameter by parameter tests. This type of parameter by parameter analysis is often necessary to discover

where in the measurement model noninvariance exists. If items are polytomous, the number and complexity of the statistical test increases drastically.

Though there is some debate about the best measures of fit, many researchers use a chi-square difference test. This difference in chi-square statistics or log-likelihood is chi-square distributed, with degrees of freedom that represent the change in the number of estimated parameters across the two models. However, with non-normal data, this procedure is more complicated because the distributional assumptions may no longer hold. Satorra and Bentler (2001) propose using their scaled chi-square as an alternative. To conduct this test, one must apply a scaling correction factor to the chi-square statistic. This is provided in statistical software, such as Mplus, when the user specifies ordered categorical items. Still, one must use this scaling factor to calculate an alternate test by hand, following guidelines specific to the statistical software they are using (Bryant & Satorra, 2012). When items are polytomous, one may need to conduct a chi-square difference test for each, individual threshold, potentially carrying out hundreds of statistical tests.

Further, if more than two or three groups are involved, this process needs to be repeated for each pair of groups individually. This results in a myriad of complicated, statistical tests where errors of inference (e.g., Type I, Type II, over-fitting) are likely. Iterative model re-specification is completely data driven, and this process capitalizes on chance, such that findings are unlikely to generalize to other samples (MacCallum et al., 1992). Thousands of statistical tests, which could easily be needed, make it difficult to arrive at the true, final model (Asparouhov & Muthén, 2014).

Partial invariance. Even if one avoids making errors of inference or capitalizing on chance, they may arrive at a model where some parameters are invariant, and others are not (i.e.,

partial invariance, Byrne, Shavelson, & Muthén, 1989). This poses a challenge because the ramifications of specifying partial invariance are unclear. Though there is a vast body of methodological research on approaches for detecting noninvariance, as has been discussed above, there is surprisingly little research on the consequences of specifying partial invariance or guidance on how to handle it (Finch & French, 2012; Schmitt & Kuljanin, 2008).

I hypothesize two primary reasons for this. First, the largest body of work in the area of invariance testing/DIF detection comes from the educational and career testing fields. In these areas, the primary focus of noninvariance testing is to uncover bias in items across gender or race. If items function differently for different groups, it becomes an issue of fairness in testing, which is discussed at length in the Standards for Educational and Psychological Testing (2014). Additionally, testing companies and certification boards risk facing media scrutiny or even legal action if items are found to function differently for minority or disadvantaged groups. The ramifications of DIF in educational or career testing are high stakes and have grave implications for social justice and equality. Thus, it seems fitting that research has focused on detecting DIF. Once DIF is detected, those items are typically removed, to ensure the highest level of test fairness. In this context, moving forward with a model where there is partial measurement invariance is not desired or relevant.

An area where partial invariance is relevant is in psychological/sociological research or noncognitive assessment, the focus of this dissertation. Though research in these areas is crucial, the conclusions are generally less high stakes for the individual person under study and the focus is on making valid comparisons at the group level. Additionally, researchers may have few items on their scale, so excluding noninvariant items might not be possible if trying to maintain adequate measurement of the construct. However, even though measurement invariance is more

relevant in psychological research, the field, as a whole, does not undertake tests of measurement invariance as standard procedure[1], like the testing fields. This is noted in numerous reviews within organizational psychology (Neal Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). Further, the practice of testing for measurement noninvariance, as a whole, is newer to psychology than it is to the testing fields. This is my second hypothesized reason as to why research on partial measurement invariance is lacking.

In searching the literature for simulation studies investigating partial measurement invariance I found a small body of work. There are generally two types of investigations: those in which the effects of noninvariance (e.g., power or bias) are studied in a multiple group model and those where a novel way of measuring the impact of noninvariance is proposed.

In considering how noninvariance impacts the estimates in a multiple group model, numerous studies have found nontrivial results. A finding shared across numerous studies is that including noninvariant items when comparing group means or variances causes errors of inference, with a greater likelihood of concluding groups are different when they are not (Chen, 2008; Finch & French, 2012; Fleishman, Spector, & Altman, 2002; Jones & Gallo, 2002; Steinmetz, 2013). There is some indication that intercept noninvariance causes more bias in mean scores (Steinmetz, 2013), whereas loading noninvariance causes more bias in mean variances (Finch & French, 2012). In contrast, Schmitt and colleagues (2011; 2008) have shown through real data applications that measurement noninvariance has little impact on structural coefficients used in subsequent analysis.

---

[1] Reasons for this are debatable, in additional to measurement invariance practice being relatively new to these types of fields, it may also be that the methodology is not accessible to a non-methodologist.

Relatedly, there have been a few attempts to quantify the impact of noninvariance on substantive research questions. Millsap and Kwok (2004) propose considering how the selection of individuals based on factor scores changes as a result of partial invariance. They argue for examining changes in selection as a way of assessing the practical impact of measurement noninvariance. They propose that researchers investigate how the individuals selected from the top quartile of the construct being measured shift under various degrees of partial invariance. They do not provide a recommendation for what should be a significant change in selection, but recommend that researchers consider this in the context of their study. Another approach, suggested by Oberski (2014), is to consider how the structural parameters in a subsequent analysis with the measurement model in question change as a result of noninvariance. Oberski describes an R-program that quantifies the expected change in the parameter of interest (EPC-interest). He offers evidence that this metric provides researchers with a substantively meaningful way to assess whether noninvariance causes problems for answering research questions.

The general dearth of research in this area may be the reason for a lack of consensus in the field regarding how to proceed under cases of partial measurement invariance. Some scholars argue that factors are comparable under partial invariance (Byrne et al., 1989; Reise et al., 1993; Steenkamp & Baumgartner, 1998), whereas others disagree (Ferrer & Widaman, 2008; Millsap & Kwok, 2004), and some simply state that there is not enough research to be certain (Neal Schmitt & Kuljanin, 2008). Ultimately, if partial invariance is found, the researcher is left to ponder how to proceed. Millsap and Meredith (2004) describe four options and their corresponding hitches: (1) go forward with the partially invariant model, ignoring the differences (2) create a cut-offs for differences and only use the scale if it meets that criterion, even though

such a cut-off would be arbitrary (3) revise the scale, only retaining the invariant items, which would create many versions of the scale, or (4) give up and do not use the scale. Unfortunately, given these challenges and complications with traditional measurement invariance testing, the task of making valid group comparisons across many schools or programs when scale items are polytomous seems insurmountable.

Methodologies for Studying Invariance across Many Groups

To date, two alternative frameworks have been proposed for investigating measurement invariance across many groups: multilevel models and the alignment method. Despite the potential to solve the same problem, the frameworks are vastly different in their theoretical and practical implications. Multilevel modeling is a random effects approach for investigating measurement invariance where item parameters are modeled as randomly varying in the population from which the groups were drawn. The alignment method, however, is a fixed effects approach, where each group is treated as a separate entity and has a separate, estimated measurement model.

Testing for measurement noninvariance in a multilevel model takes on a different form than is seen in other approaches. The model is specified like a CFA, in that items are mathematically linked to a composite factor. However, in a MLM framework, one can specify item intercepts and loadings as randomly varying across a population. It is the variance of the item parameters across groups that becomes of interest when considering measurement noninvariance. De Jong, Steenkamp, and Fox (2007) propose fixing level two variances of measurement parameters and comparing model fit as a test of noninvariance. If model fit is compromised, or if the variance of these parameters is significant, it is considered evidence of noninvariance. Muthén and Asparohouv (2013b) suggest a slightly different approach, allowing

random, yet small variation (e.g., .01) in item parameters across groups. Though both solutions have been shown viable in deriving comparable factor means, more research is needed to understand the limitations and implications.

It is worth noting that the use of MLMs for investigating invariance across many groups, or all level two units, is different than more traditional DIF testing in an MLM framework. Investigating differences across a few groups, such as races or gender, is commonly undertaken in MLMs to account for nesting of students within schools. This has been discussed in depth (e.g., Beretvas, Cawthon, Lockhart, & Kaye, 2012; Beretvas & Walker, 2012; French & Finch, 2010) and typically involves using categorical predictors of group membership to explain variability in item parameters across groups. For example, if gender significantly predicts variance in item intercepts across groups, such that the residual variance is non-significant (i.e., non-randomly varying intercepts) one could conclude DIF across genders. Of course, these predictors could be used in tandem with investigations of the residual variance to undercover person and situational factors that influence the differences in item parameters across many groups.

In further considering the contrast between MLMs and the alignment method, there are a number of important assumptions. First, MLMs have the assumption that item parameters are random from normally distributed populations. However, this assumption may not be tenable in many situations. Muthén and Asparohouv (2013) discuss an example with country-level data, where it is reasonable to view groups as separate, fixed entities. It is likely with country-level data that just a few groups are quite different from the rest. This noninvariance could result in non-normal variance of the level-two item parameters, a violation of that assumption. This

situation also seems reasonable when comparing American states or schools. The alignment method does not assume that the groups are from the same population, but are separate entities.

Second, there are critical sample size considerations for MLMs and the alignment method. MLMs require a large number of level two units, upwards of 50, for unbiased estimates of level two variances (Maas & Hox, 2005). This is a key assumption when testing for measurement invariance because the variances of level-two item parameters are of primary interest. In contrast, for the alignment method, the within-group sample size is of more importance, and there is not a minimum number of required groups. The alignment method has the potential to fill a unique space where researchers have a number of large groups, too many to do a traditional multiple group CFA, but not enough to satisfy the assumptions of a multilevel model. Consequently, the alignment method may not be well suited for situations where the researcher has a large number of small groups, which is an ideal situation for the using multi-level models.

Muthén and Asparohouv (2013) provide a comprehensive comparison of the two approaches, which pinpoints caveats for using one or the other. However, there are still many unanswered questions when considering the two methods for assessing measurement invariance and how to best choose between them. This dissertation will focus on evaluating the alignment method for tests of noninvariance across many groups. The alignment method is a new technique and more research is needed to understand the assumptions and ideal circumstances for using the method, particularly with polytomous items.

Overview of the Alignment Method

The alignment method (Asparouhov & Muthén, 2014) was recently proposed as an alternative means of estimating confirmatory factor models across many groups. This method

was proposed as a way to simplify the unwieldy task of multiple-group measurement. The method automates the estimation of model parameters across many groups and testing for noninvariance. This method is carried out in a completely different fashion than the traditional tests I described above. It works in two parts: first a simplicity function similar to the rotation criteria executed in exploratory factor analysis is used to estimate group specific measurement models. The measurement model parameters are estimated in a way that minimizes noninvariance, such that factor means and variances are comparable across groups. After estimation of the group specific measurement models, an ad-hoc testing procedure is used to identify substantial noninvariance of model parameters.

The Alignment Optimization Procedure: Estimating Measurement Models across Many Groups

Asparohouv and Muthén (2014) describe the estimation procedure employed in the alignment. They begin with considering the multiple group factor model for continuous items:

$$y_{ijg} = \nu_{jg} + \sum_{l=1}^{L} \lambda_{jlg} \eta_{ilg} + \varepsilon_{ijg} \qquad (13)$$

Where $i = 1\ldots, I$, and I is the number of people, $j = 1\ldots, J$, and J is the number of items, $g = 1, \ldots, G$ and G is the number of groups. η_{ilg} is the latent variable for person i, in group g, on factor l and ε_{ijg} is the residual, as described above. The alignment optimization begins with the estimation of a model where the factor mean, $\alpha = 1$, and the factor variance, $\psi = 1$ for every group, g. This is the base model, M0, and all intercepts (or thresholds) and loadings are estimated as free and unequal (i.e., a configural model). This configural model, M0, transforms the factor in each group to mean zero and variance one:

$$\eta_{g0} = (\eta_g - \alpha_g) / \sqrt{\psi_g} \qquad (14)$$

This transformation allows the variance and mean of the indicators to be expressed as:

$$\begin{aligned} V(y_{jg}) &= \lambda_{jg}^2 \Psi_g = \lambda_{jg,0,}^2 \\ E(y_{jg}) &= \nu_{jg} + \lambda_{jg}\alpha_g = \nu_{jg,0} \end{aligned} \tag{15}$$

where estimates of the base model, M0, are denoted $\nu_{jg,0}$, and $\lambda_{jg,0}$ and

$$\begin{aligned} \lambda_{jg,0} &= \lambda_{jg}\sqrt{\psi_{g,}} \\ \nu_{jg,0} &= \nu_{jg} + \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}\alpha_g \end{aligned} \tag{16}$$

For every set of group factor means and variances there are intercept and loading parameters that yield the same likelihood as the configural model. You can obtain these estimates using the following equations:

$$\begin{aligned} \lambda_{jg,1} &= \frac{\lambda_{jg,0}}{\sqrt{\psi_g}} \\ \nu_{jg,1} &= \nu_{jg,0} - \alpha\frac{\lambda_{pg,0}}{\sqrt{\psi_g}}. \end{aligned} \tag{17}$$

The intercepts and loadings are chosen such that the measurement noninvariance between the groups is minimized. This is done with respect to α_g and ψ_g using a loss/simplicity function, F, that accumulates the total measurement noninvariance:

$$\begin{aligned} F = &\sum_j \sum_{g_1<g_2} w_{g_1,g_2} f(\lambda_{jpg_{1,}1} - \lambda_{jpg_{2,}1}) \\ &+ \sum_j \sum_{g_1<g_2} w_{g_1,g_2} f(\nu_{jpg_{1,}1} - \nu_{jpg_{2,}1}) \end{aligned}. \tag{18}$$

For every pair of groups' intercepts and loadings the difference between the parameters are accumulated and then scaled by *f*, the component loss function (CLF). The groups are weighted by their sample size, *w*, such that larger groups will contribute more to the total loss function, *F*. The CLF has also been used in EFA rotation algorithms to estimate factor loadings with the simplest possible structure (Jennrich, 2006). Asparohouv and Muthén discuss the base model,

M0, and the resulting model M1 as paralleling the unrotated and rotated factor solutions in EFA. In the alignment the CLF is:

$$f(x) = \sqrt{\sqrt{x^2 + \varepsilon}}\ . \tag{19}$$

In the alignment optimization procedure ε is a small number, .01. A positive ε is used so that the CLF has a continuous first derivative, simplifying optimization of the total loss function, F. The total loss function is minimized at a solution where there are few large noninvariant parameters and many approximately invariant parameters. The function does not optimize when there are many medium sized noninvariant parameters. Thus, there is an assumption that a majority of the parameters are approximately invariant, with only a few substantially noninvariant ones. Again, Asparohouv and Muthén compare this to EFA, in that rotation functions attempt to simplify the loading matrix with either large or small loadings, not midsized ones. However, there are not currently any rules of thumb to guide the researcher in what constitutes approximate invariance or large noninvariance between groups.

Upon minimizing the total loss function, 2*G*-1 of the parameters in the model can be estimated. Identification is achieved by estimating all groups' factor means and variances except the first group, where the following constraint is used:

$$\psi_1 \times ... \times \psi_g = 1 \tag{20}$$

However, the first group's factor mean can also be constrained to zero. This produces two identification options in the alignment, FREE (only ψ_1 is constrained) and FIXED (α_1 and ψ_1 are constrained). The optimization procedure is conducted after all of the indicators are standardized across the whole population, such that all variables are on the same scale and the loss functions between the different indicators are comparable. The parameters are reported in a standardized

metric (i.e., ψ_1=1). The procedure first obtains the factor means and variances for all the groups, then solves for the intercepts and loadings parameters via the equations above.

The Alignment Ad-hoc Testing Procedure

Though the primary purpose of the alignment is to produce comparable factor means and variances, the procedure does produce output about the degree of measurement invariance. After the group specific measurement models are estimated, invariance testing is conducted on all of the parameters. Taking one parameter at a time, two groups are compared. If these groups are not statistically significantly different from one another they become connected. These comparisons are made again and again, across the groups to create an invariant set of groups, and then each parameter is tested against the mean of the invariant set. If that parameter, for that group is significantly different from the mean, then it is flagged as a noninvariant parameter. Asparohouv and Muthén control for Type I error rate in the algorithm by setting the alpha value to .001. The program produces, for each parameter in the model, which groups are invariant, the mean differences between every pair of groups in the analysis, as well as the corresponding p-value for the pairwise differences.

The ad-hoc output also includes an R^2 measure that describes the variability explained in the measurement parameters across groups that is due to group mean and variances differences (Asparouhov & Muthén, 2014). An R^2 near 1 indicates complete invariance because the variability in item parameters are completely explained by group mean differences, whereas an R^2 near zero indicates that group mean differences explain none of the variability in item parameters. This metric, in addition to the significance testing provides information about the magnitude of noninvariance in an item across groups.

Alignment with Ordered Categorical Items

In 2013 when the alignment method was introduced as a part of MPlus version 7.1, factor models with continuous and binary items could be estimated. Polytomous item capability was introduced in 2014 as a part of MPlus version 7.3. Muthén and Asparouhov (2014) discuss the alignment in the specific context of ordered categorical items. The alignment functions in the same way as the ordered categorical factor model discussed previously, in that the observed score, y_{ijg} is assumed to be a continuous latent response variable $^*y_{ijg}$, partitioned by thresholds, τ. Further, the alignment method is carried out in the THETA parameterization, which results in thresholds and loadings that are in probability units.

Performance of the Alignment Method

Asparohouv and Muthén (2014) presented results from the first set of simulation studies investigating the performance of the alignment method. They report on three simulation studies: (1) recovery and bias of model parameters with maximum likelihood (ML) estimation, (2) comparison of Bayesian and maximum likelihood estimators, and (3) comparison of the FREE and FIXED identification options within the alignment. In their first simulation they generate a five indicator, single-factor model where indicators are continuous, ranging from approximately -5 to 5. The conditions of the study are summarized in Table 1.

Table 1.
Asparohouv and Muthén (2014) Simulation Conditions

Manipulated Factors	Conditions
Number of groups	2,3,15,60
Within group n	100, 1000
Amount of noninvariance	0%, 10%, 20%
Identification options	Fixed or Free

The groups were generated such that they follow a pattern, where three group types were repeated to equal the total number of groups. The values for each measurement model parameter by group type are shown in Table 2. The measurement parameters are listed in the first column, then in the columns two, three, and four each group's specific value for the parameter is listed. For example, the intercept for Y1 was set to zero in groups one and three, and group two was noninvariant with an intercept of -.5. This general pattern of noninvariance was consistent across conditions. The amount of noninvariance was in regards to the number of measurement parameters within a group that were noninvariant. For example, in Table 2, 1 intercept and 1 slope are noninvariant, or 10% of the intercepts and slopes.

Table 2.
Noninvariance Pattern in Asparouhov and Muthén (2014)

Parameter	Group 1	Group 2	Group 3	Number of Groups N.I.
Y1 Int	0	-.5	0	1/3
Y2 Int	0	0	.5	1/3
Y3 Int	0	0	0	0
Y4 Int	0	0	0	0
Y5 Int	.5	0	0	1/3
Y1 Slope	1	1	1	0
Y2 Slope	1	1	1	0
Y3 Slope	1.4	1	1	1/3
Y4 Slope	1	1	.3	1/3
Y5 Slope	1	.5	1	1/3
Factor Mean	0	.3	1	-
Factor Variance	1	1.5	1.2	-

Note. Residual Variances = 1

Muthén and Asparouhov (2014) discuss the alignment method particularly in the context of IRT models with binary items and provide further illustration. They suggest a rough recommendation for the alignment method: a limit of 25% noninvariance for trustworthy alignment results. However, they suggest conducting a simulation study using the real data as starting values to evaluate the appropriateness of the alignment method for a given dataset. In evaluating the simulation from real data, Muthén and Asparouhov suggest using the correlation

between true factor means and estimated factor means as a gauge of the performance of the alignment. They state that true and estimated factor mean correlations of .98 were sufficient in producing little bias in the group specific factor means.

This series of simulation studies provided evidence that the alignment method can accurately recover parameter estimates for many groups when group sizes are large and there is little or no noninvariance. As noninvariance increased and sample sizes decreased there was bias in the estimates. For the FIXED identification option, absolute bias in factor means, factor variances, intercepts, and loadings was only greater than .05 when noninvariance was large (20% of parameters) or when there were a large number of groups, $n = 60$, and a small within group sample size, $n = 100$. In all of the large group size conditions, $n = 1000$, bias was less than .05 across all parameters in the model. This was expected with maximum likelihood estimation, which is reliant on large samples.

The FREE identification option showed greater bias overall across all of the parameters in the model. Absolute bias was greater than .05 in most of the small within group *n* conditions. Further, there was substantial bias in the larger sample size conditions when there were only two groups. Generally it seems that the FREE options works best when there is a moderate to large amount of noninvariance (10-20% of parameters) and there are more than two groups. Thus, it seems reasonable that the FREE option is preferred when the researchers have many groups and suspect some noninvariance.

This simulation work provides great promise for use in ascertaining measurement invariance of polytomous scales across many groups. The purpose of the dissertation research is to build upon this prior work by learning more about how partial measurement invariance, a violation of the assumption of minimal noninvariance, manifests in the alignment results. I will

extend previous research by evaluating the methodology under four new conditions: (1) realistic, Likert scale data from a large scale noncognitive assessment, (2) a two-factor model, and (3) manipulating the degree of group differences. Further, I will conduct an investigation of the ad-hoc testing procedure, which was not presented in Asparohouv and Muthén (2014).

Chapter 3: Methodology

This study took place in three parts: real data analysis, simulation, and output analysis. First, real item response data were analyzed to obtain realistic specifications for data generation. After obtaining the population parameters for the simulated non-cognitive scales, I manipulated those parameters to meet various conditions. Five-hundred replications of each condition were carried out with the alignment method. Finally, I analyzed the resulting output from both parts of the alignment: group specific measurement models and testing procedure.

Real Data Analysis

Existing, secure data from a large testing program were analyzed to obtain realistic starting values for the simulation. The data are from a non-cognitive battery of assessments, designed for use with middle school students, that has been collected at schools all across the United States. The data are from the fall 2013 administration and include 12,719 complete responses. Ordinal, confirmatory factor analyses were conducted with two of the subscales, time management and intrinsic motivation. Each subscale includes seven items describing a behavior or attitude, e.g., "I love to learn" with a four point response scale ranging from 1, "never or rarely", to 4, "usually or always."

Factor models were run in Mplus version 7.2 using the THETA parameterization. Since the alignment is carried out in the THETA parameterization, it was important that the simulation generating values are also in that metric. A 2-factor model fit the data adequately. The chi-square test of model fit was significant, χ^2 (76) = 4453.01, but all other indices were acceptable: RMSEA= .067, CFI=.971, and TLI .965. The factors were moderately correlated, r =.56. Estimates from this single group, 2-factor model were used to create the overall population measurement model in the simulation.

Simulation Study

The overall population measurement model was manipulated to fit various conditions. The simulation was carried out in Mplus version 7.2, using the Monte Carlo features available in the program. The simulation study inputs were modeled using examples provided on statmodel.com by the authors of the software. 500 replications from each condition were completed as suggested by Muthén (2002). The following is an example annotated input file used in the simulation (more example inputs are included in appendix A):

```
MONTECARLO:
   NAMES = y1-y14;
   GENERATE= y1-y14 (3);
   !setting response scale of variables;
   CATEGORICAL = y1-y14;
   !specifying categorical;
   NOBSERVATIONS = 3(500);
   !number of observations per group;
   NGROUPS = 3;
   !number of groups;
   NREPS = 500;
   !number of replications

ANALYSIS:
   TYPE =MIXTURE;
   ESTIMATOR = MLR;
   !specifying maximum likelihood robust;
   alignment = fixed;
   !specifying the fixed identification option;
   ALGORITHM=INTEGRATION;
   processors=8;

MODEL POPULATION:
    %OVERALL%
        TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
         y5*1.006 y6*-.536 y7*.904;
    !specifying the overall model that is common across groups;
    !specifying the factor and factor loadings;

    [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
    [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
    [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
    [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
    [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
    [y6$1*.661 y6$2*1.87 y6$3*2.42];
    [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];
    !specifying the thresholds;
      IM by y8*1.715 y9*1.391 y10*.953 y11*.897
```

```
          y12*1.253 y13*1.281 y14*1.571;
    !second factor and loadings;
        [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
    [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
    [y10$1*-.706 y10$2*.85 y10$3*1.892];
    [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
     [y12$1*-3.025 y12$2*-.996 y12$3*.831];
    [y13$1*-2.855 y13$2*-.716 y13$3*.844];
    [y14$1*-3.487 y14$2*-1.252 y14$3*.878];
  !second factor thresholds;
  !below are the group specific starting values;
  %G#1%
    [tm*0];
    !factor mean is 0;
        tm*1;
     !factor variance is 1;
     [im*0];
     im*1;
    tm with im*.56;
      !factor covariance;

              TM by Y1*.709;

              IM by Y8*1.315;

              !noninvariance on loadings for one item on each scale;

  %G#2%
  [tm*0.3];
        tm*1.5;
   [im*0.3];
        im*1.5;
   tm with im*.83;

          TM by Y3*.887;

          IM by Y10*.553;

              !noninvariance for one item on each scale;

  %G#3%
   [tm*1];
        tm*1.2;
   [im*1];
   im*1.2;
   tm with im*.67;

          TM by Y5*.606;

          IM by Y12*.853;

              !noninvariance for one item on each scale;
model:

  %OVERALL%
        TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
```

```
        y5*1.006 y6*-.536 y7*.904;

  [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
  [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
  [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
  [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
  [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
  [y6$1*.661 y6$2*1.87 y6$3*2.42];
  [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

    IM by y8*1.715 y9*1.391 y10*.953 y11*.897
         y12*1.253 y13*1.281 y14*1.571;

        [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
  [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
  [y10$1*-.706 y10$2*.85 y10$3*1.892];
  [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
   [y12$1*-3.025 y12$2*-.996 y12$3*.831];
  [y13$1*-2.855 y13$2*-.716 y13$3*.844];
  [y14$1*-3.487 y14$2*-1.252 y14$3*.878];

 %G#1%
  [tm*0];
  !factor mean is 0;
        tm*1;
    !factor variance is 1;
   [im*0];
   im*1;
  tm with im*.56;
    !factor covariance;

                TM by Y1*.709;

                IM by Y8*1.315;
                !noninvariance on loadings for one item on each scale;

  %G#2%
  [tm*0.3];
        tm*1.5;
   [im*0.3];
        im*1.5;
   tm with im*.83;

                TM by Y3*.887;

                IM by Y10*.553;

                !noninvariance for one item on each scale;

  %G#3%
   [tm*1];
        tm*1.2;
   [im*1];
   im*1.2;
   tm with im*.67;
```

```
        TM by Y5*.606;
        IM by Y12*.853;
            !noninvariance for one item on each scale;
output: align;
```

The input begins with the MONTECARLO statement, which includes the specifications for data generation. The second part of the input includes the analysis options where the estimator and identification options for the alignment are called. Under the MODEL POPULATION command are the specifications for the data generation model. I first list the overall specifications and then the individual group specifications using the %G#% command, if they are different from the overall parameters. After specifying the population model, one then has to specify the analysis model. As in the alignment simulation conducted by Asparohouv and Muthén (2014), I specified the model correctly and provided correct starting values. Finally, the align output is requested, which produces the testing procedure information for each replication.

The conditions of the simulation study are summarized in Table 3. The sample size, pattern of noninvariance, factor model, scale length, estimator, and identification option were constant factors. The number of groups, percentage of noninvariance, magnitude of noninvariance, and type of noninvariance were manipulated factors. Excluding the 0% level of noninvariance, all conditions were crossed, totaling 57 conditions. Figure 2 shows the conditions in more detail for the 15 groups set of conditions. The manipulated factors are discussed in more detail below.

Table 3.
Simulation Study Factor Descriptions

Manipulated Factors	
Number of groups	3, 9, 15
Number of noninvariant item parameters	0%, 14%, 29%, 43%
Magnitude of noninvariance	Small, medium, large
Type of noninvariance	Loading or threshold
Constant Factors	
Pattern of noninvariance	See Tables 4-6.
Sample size	500 per group
Factor model	2 correlated factors
Scale length	14 items total, 7 on each factor
Estimator	ML
Identification option	FIXED

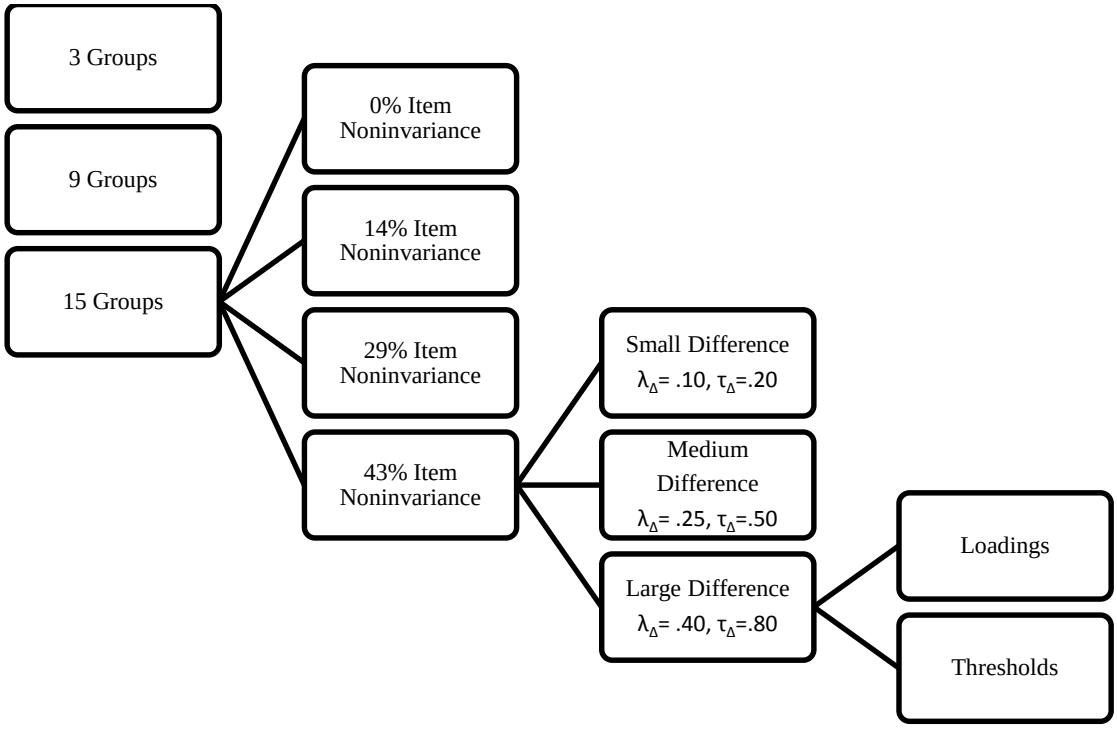

Figure 2. Conditions of Simulation

Pattern of noninvariance. Utilizing the same method as Asparohouv and Muthén (2014), I generated three group types, and then repeated those group types to create the 9 and 15 groups conditions. Each group type has the same factor mean, factor variance, and pattern of noninvariant items. For group type 1, $\alpha = 0$, $\psi = 1$, group type 2, $\alpha = .3$, $\psi = 1.5$, group type 3, $\alpha = 1$, $\psi = 1.2$, which is consistent with the group types used in Asparohouv and Muthén (2014). For example, in the 9 group conditions, the first, fourth, and seventh groups are simulated in the same manner. Tables 4-6 include the pattern of loading and threshold noninvariance by group for the small magnitude conditions. The items with simulated noninvariance are shaded in the table. The medium and large magnitude conditions were simulated in the same manner, the starting values for those sets of conditions are included in Appendix B.

Table 4.

Simulated Loading Values, Pattern of Noninvariance, for Small Magnitude Conditions, Loadings -.10 from Overall

Overall, 0% N.I. Items			14% N.I. Items			29% N.I. Items			43% N.I. Items [a]		
Subscale	Item	Loading	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
TM	Y1	1.109	1.009	1.109	1.109	1.009	1.109	1.109	1.009	1.109	1.109
TM	Y2	0.950	0.95	0.95	0.95	0.85	0.95	0.95	0.85	0.95	0.95
TM	Y3	1.287	1.287	1.187	1.287	1.287	1.187	1.287	1.287	1.187	1.287
TM	Y4	1.010	1.01	1.01	1.01	1.01	0.91	1.01	1.01	0.91	1.01
TM	Y5	1.006	1.006	1.006	0.906	1.006	1.006	0.906	1.106	1.006	0.906
TM	Y6	-0.536	-0.536	-0.536	-0.536	-0.536	-0.536	-0.436	-0.536	-0.536	-0.436
TM	Y7	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.804
IM	Y8	1.715	1.615	1.715	1.715	1.615	1.715	1.715	1.615	1.715	1.715
IM	Y9	1.391	1.391	1.391	1.391	1.291	1.391	1.391	1.291	1.391	1.391
IM	Y10	0.953	0.953	0.853	0.953	0.953	0.853	0.953	0.953	0.853	0.953
IM	Y11	0.897	0.897	0.897	0.897	0.897	0.797	0.897	0.897	0.797	0.897
IM	Y12	1.253	1.253	1.253	1.153	1.253	1.253	1.153	1.353	1.253	1.153
IM	Y13	1.281	1.281	1.281	1.281	1.281	1.281	1.181	1.281	1.281	1.181
IM	Y14	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.471

Note. Y6 is a negatively worded item, for this item only the N.I. was simulated as +.10 instead of -.10, resulting in a poorer loading for that item.

[a] In the 43% N.I. condition, Y5 and Y12 are different across all three groups and no item is completely invariant.

Table 5.

Simulated Time Management Threshold Values, Pattern of Noninvariance, for Small Magnitude Conditions, Thresholds -.20 from Overall

Overall, 0% N.I. Items			14% N.I. Items			29% N.I. Items			43% N.I. Items [a]		
Subscale	Parameter	Threshold	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
TM	Y1$1	-3.617	-3.817	-3.617	-3.617	-3.817	-3.617	-3.617	-3.817	-3.617	-3.617
TM	Y1$2	-2.153	-2.353	-2.153	-2.153	-2.353	-2.153	-2.153	-2.353	-2.153	-2.153
TM	Y1$3	-0.541	-0.741	-0.541	-0.541	-0.741	-0.541	-0.541	-0.741	-0.541	-0.541
TM	Y2$1	-1.776	-1.776	-1.776	-1.776	-1.976	-1.776	-1.776	-1.976	-1.776	-1.776
TM	Y2$2	-0.481	-0.481	-0.481	-0.481	-0.681	-0.481	-0.481	-0.681	-0.481	-0.481
TM	Y2$3	0.67	0.67	0.67	0.67	0.47	0.67	0.67	0.47	0.67	0.67
TM	Y3$1	-2.796	-2.796	-2.996	-2.796	-2.796	-2.996	-2.796	-2.796	-2.996	-2.796
TM	Y3$2	-1.141	-1.141	-1.341	-1.141	-1.141	-1.341	-1.141	-1.141	-1.341	-1.141
TM	Y3$3	0.432	0.432	0.232	0.432	0.432	0.232	0.432	0.432	0.232	0.432
TM	Y4$1	-1.92	-1.92	-1.92	-1.92	-1.92	-2.12	-1.92	-1.92	-2.12	-1.92
TM	Y4$2	-0.311	-0.311	-0.311	-0.311	-0.311	-0.511	-0.311	-0.311	-0.511	-0.311
TM	Y4$3	1.036	1.036	1.036	1.036	1.036	0.836	1.036	1.036	0.836	1.036
TM	Y5$1	-3.284	-3.284	-3.284	-3.484	-3.284	-3.284	-3.484	-3.084	-3.284	-3.484
TM	Y5$2	-2.066	-2.066	-2.066	-2.266	-2.066	-2.066	-2.266	-1.866	-2.066	-2.266
TM	Y5$3	-0.597	-0.597	-0.597	-0.797	-0.597	-0.597	-0.797	-0.397	-0.597	-0.797
TM	Y6$1	0.661	0.661	0.661	0.661	0.661	0.661	0.461	0.661	0.661	0.461
TM	Y6$2	1.87	1.87	1.87	1.87	1.87	1.87	1.67	1.87	1.87	1.67
TM	Y6$3	2.42	2.42	2.42	2.42	2.42	2.42	2.22	2.42	2.42	2.22
TM	Y7$1	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-3.139
TM	Y7$2	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.733
TM	Y7$3	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.277

[a] In the 43% N.I. condition, Y5 thresholds are different across all three groups and no item is completely invariant.

Table 6.
Simulated Intrinsic Motivation Threshold Values, Pattern of Noninvariance, for Small Magnitude Conditions, Thresholds -.20 from Overall

Overall, 0% N.I. Items			14% N.I. Items			29% N.I. Items			43% N.I. Items [a]		
Subscale	Parameter	Threshold	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
IM	Y8$1	-3.152	-3.352	-3.152	-3.152	-3.352	-3.152	-3.152	-3.352	-3.152	-3.152
IM	Y8$2	-0.768	-0.968	-0.768	-0.768	-0.968	-0.768	-0.768	-0.968	-0.768	-0.768
IM	Y8$3	1.038	0.838	1.038	1.038	0.838	1.038	1.038	0.838	1.038	1.038
IM	Y9$1	-2.957	-2.957	-2.957	-2.957	-3.157	-2.957	-2.957	-3.157	-2.957	-2.957
IM	Y9$2	-0.776	-0.776	-0.776	-0.776	-0.976	-0.776	-0.776	-0.976	-0.776	-0.776
IM	Y9$3	1.09	1.09	1.09	1.09	0.89	1.09	1.09	0.89	1.09	1.09
IM	Y10$1	-0.706	-0.706	-0.906	-0.706	-0.706	-0.906	-0.706	-0.706	-0.906	-0.706
IM	Y10$2	0.85	0.85	0.65	0.85	0.85	0.65	0.85	0.85	0.65	0.85
IM	Y10$3	1.892	1.892	1.692	1.892	1.892	1.692	1.892	1.892	1.692	1.892
IM	Y11$1	-2.656	-2.656	-2.656	-2.656	-2.656	-2.856	-2.656	-2.656	-2.856	-2.656
IM	Y11$2	-1.075	-1.075	-1.075	-1.075	-1.075	-1.275	-1.075	-1.075	-1.275	-1.075
IM	Y11$3	0.31	0.31	0.31	0.31	0.31	0.11	0.31	0.31	0.11	0.31
IM	Y12$1	-3.025	-3.025	-3.025	-3.225	-3.025	-3.025	-3.225	-2.825	-3.025	-3.225
IM	Y12$2	-0.996	-0.996	-0.996	-1.196	-0.996	-0.996	-1.196	-0.796	-0.996	-1.196
IM	Y12$3	0.831	0.831	0.831	0.631	0.831	0.831	0.631	1.031	0.831	0.631
IM	Y13$1	-2.855	-2.855	-2.855	-2.855	-2.855	-2.855	-3.055	-2.855	-2.855	-3.055
IM	Y13$2	-0.716	-0.716	-0.716	-0.716	-0.716	-0.716	-0.916	-0.716	-0.716	-0.916
IM	Y13$3	0.844	0.844	0.844	0.844	0.844	0.844	0.644	0.844	0.844	0.644
IM	Y14$1	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.687
IM	Y14$2	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.452
IM	Y14$3	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.678

[a] In the 43% N.I. condition, Y12 thresholds are different across all three groups and no item is completely invariant.

Number of groups. The three groups condition was chosen because it is still very tedious to conduct a multiple group CFA with three groups, thus if the alignment works well it provides an alternative to traditional testing procedures. Also, three groups are likely if educators are interested in differences across school type, major field of study, or demographics. The 9 and 15 groups conditions, though still too many groups to do a traditional multiple-group CFA, are not enough to use an MLM framework. These numbers of groups represent situations in which the alignment might be ideal.

Amount of noninvariance. Asparohouv and Muthén reported little to no bias when noninvariance was exhibited in 0% or 10% of the parameters. Even with 20% noninvariance, only small sample conditions showed absolute bias greater than .10. I investigated the method with an increased amount of noninvariance to better discover where issues might arise. I generated uniform noninvariance on 0, 1, 2, and 3 items on each subscale, which corresponds to 0, 14, 29, and 43 percent of items. The 14% and 29% conditions are similar to those simulated in Finch and French (2008). They describe 17% noninvariant loadings as low contamination and 34% as high contamination. The 43% of noninvariant items condition represents a more extreme situation. In these conditions there is one item per subscale that is completely noninvariant, with simulated differences across all three group types. Also in the 43% noninvariance conditions, no item is completely invariant across all three group types. This can be seen in tables 4-6 and is described in the table notes.

Magnitude of noninvariance. The testing procedure that accompanies the alignment optimization is not investigated by Asparohouv and Muthén (2014), thus there is much to be learned about how well it functions under various conditions of noninvariance. In varying the magnitude of group differences, I could systematically investigate how large group differences

must be to be flagged as noninvariant by the procedure. The simulated differences in loadings were: small = -.10, medium = -.25, and large = -.40. For the thresholds the differences were: small = -.20, medium = -.50, and large = -.80. As discussed above, the data generation model was executed in Mplus's THETA parameterization, to match the estimation method of the alignment; therefore these differences are in probability units and akin to estimates from an IRT normal ogive model. These differences were chosen because they are consistent with differences simulated in other measurement invariance/differential item functioning studies. For example, using the normal ogive model, Stark, Chernyshenko, and Drasgow (2006) simulated differences as .40 and .15 in loadings and .50 and .25 in thresholds as large and small respectively. Because a primary focus of this study is to examine differences in results based on magnitude of item noninvariance, we have included three levels of noninvariance magnitude.

Table 7 includes the simulated values for one item in the small and large difference conditions as well as the corresponding probability. The probabilities provide an accessible way of interpreting the magnitude of differences simulated and how they manifest in item responding. The table lists the invariant threshold of -.71 and for the difference simulated in the small conditions the noninvariant threshold becomes -.91, a difference in the threshold of .20. This manifests as a very slight decrease in the probability of transitioning to the next category from .24 for the invariant groups to .18 for the noninvariant groups. However, in the large magnitude conditions the change in probability is from .24 to .07, an arguably substantial decrease in the probability. Another useful way to think of these probit units is to approximate the corresponding logit, as many studies on measurement invariance and DIF use a logistic link function in the factor model. The approximate logit of the probit unit is obtained by multiplying the probit by 1.7. In the logit metric the differences would correspond to .17, .43, and .68 for the

small, medium, and large loading differences and .34, .85, and 1.36 for the small, medium, and large threshold differences. Again, these differences are consistent with those commonly simulated in DIF research, where differences tend range from .3 to .7 (Woods, 2009).

Table 7.
Simulated Threshold Values and Corresponding Probabilities for Item 10 in the Small Magnitude Conditions

Invariant Threshold	Corresponding Probability	Noninvariant Threshold (SM)	Corresponding Probability (SM)	Noninvariant Threshold (LG)	Corresponding Probability (LG)
-0.71	.24	-.91	.18	-1.51	.07
0.85	.80	.65	.75	0.05	.52
1.89	.97	1.70	.96	1.09	.86

Type of noninvariance. In one set of conditions there is noninvariance on the loadings, then in a separate set there is noninvariance on the thresholds. This coincides with traditional tests for invariance in SEM, in that researchers typically investigate pattern/metric invariance (i.e., invariance of the loadings) first, before testing the thresholds to achieve scalar invariance. Further this simplifies the simulation considerably. This type of simplification is often made in simulation studies of measurement invariance/DIF where either loading or threshold noninvariance is simulated (e.g., Cheung & Rensvold, 2002; French & Finch, 2008; Meade & Bauer, 2007).

Output Analyses

The output analyses include two main sets: accuracy and precision of measurement model estimates and hit rate of the ad-hoc testing procedure. These analyses were conducted on information compiled from secondary programs written in FORTRAN 90 by myself and a colleague, Erin Strauts. Generally, the programs open the output files produced by Mplus and read the necessary information from the files to calculate various summary statistics (MSE, hit

rate, etc.) The programs produce data files that were then opened in Excel. The Pivot Tables feature in Excel was used to create the summary tables and graphs.

Measurement estimates analyses. To understand how well the alignment estimates the group specific measurement models I included a variety of measures of accuracy and precision. First, I report the coverage of loadings, thresholds, factor means, and factor variances. Coverage is the percentage of times the true value of the parameter was in the 95% confidence interval across the 500 replications (Muthén, 2002) and is reported by the Mplus software. Coverage should be close to .95, coinciding with the 5% error rate of a 95% confidence interval. To summarize the coverage across different types of parameters in the model, I calculated the coverage of each parameter type by averaging across the coverage of all of the specified parameters. For example, table 8 reports the average coverage of loadings across conditions, these numbers represent the average coverage value of all the loadings across all groups from each condition, not for a specific item or group.

I also calculated the relative bias and mean square error (MSE) for different parameter types. The relative bias is calculated in the same manner as Hoogland and Boomsma (1998), where the true value is subtracted from the estimated value and that difference is divided by the true value. This puts the bias on a percentage metric, making it easier to compare across conditions where there are varying degrees of noninvariance. Hoogland and Boomsma suggested that .05 or less is an acceptable level of relative bias. The MSE captures the bias and variability of the estimates by summing the square of the bias and the variance of the estimate. As with the coverage values, I report the average relative bias and MSE for parameter type: loading, threshold, factor mean, and factor variance.

Finally, in addition to investigating the relative bias and MSE of factor means and variances, I also examined the correlation between generated group means and estimated means across replications. This information was included as a part of the alignment simulation output and was utilized in other simulations studies on the alignment (e.g., Muthén & Asparohouv, 2013) as a crude measure of factor means recovery. Though the correlation between true and estimated factor means does not reveal if there is bias, it does quantify how well the rank order of the group means was recovered.

Testing procedure. The noninvariance testing procedure that accompanies the alignment produces hundreds of pages of output, even for a single replication. Figure 3 shows an example of the output for a single item's thresholds. This information is provided for each item by parameter (intercept/threshold or loading). The user is provided with pairwise tests for each pair of groups by parameter and a list of groups that are invariant for each parameter. In the case of a simulation, this information is provided for each replication separately. To calculate the percentage of times a group was classified as invariant, Erin Strauts wrote a compilation program. This program counts the number of times the group was listed as invariant and then divides by the number of replications. This is repeated for each parameter in the model. For example, in a simulation with 2 replications, if group 1 is listed as invariant for loading 1 in replication one, but not replication 2, the group was flagged as invariant 50% of the time. From these compiled results, I calculated the hit rate, or the percentage of times the procedure listed a group as invariant for a parameter when it was, and did not list a group as invariant for a parameter when it was not.

```
REPLICATION 1:
INVARIANCE ANALYSIS

 Intercepts/Thresholds
 Threshold Y1$1
 Group     Group      Value      Value     Difference  SE        P-value
     2         1     -2.494     -2.543      0.048      0.154       0.753
     3         1     -2.545     -2.543     -0.003      0.147       0.986
     3         2     -2.545     -2.494     -0.051      0.168       0.762
 Approximate Measurement Invariance Holds For Groups:
 1 2 3
 Weighted Average Value Across Invariant Groups:      -2.528
 R-square/Explained variance/Invariance index:       0.997
```

Figure 3. Example Alignment Output

Chapter 4: Results

Convergence Rates

Convergence was achieved for every replication, across all conditions. However, some estimates, particularly group specific factor variances, were rarely too large, such that MPlus printed "*********". In such cases, I double checked the inputs for those groups and verified that there were no clerical errors. Instances where a particular group was excluded from the output analysis, for this reason, are reported as *Notes* in the corresponding tables.

Group Specific Measurement Models

Coverage. Tables 8-11 show the average coverage of the loadings, thresholds, factor means, and factor variances. The coverage was averaged across applicable parameters and groups to get a summary value. For example, when calculating the average coverage for the factor means, the first group was not included in the calculation of the average, because the factor mean is fixed to 1 and not estimated in that group. The tables show that coverage is quite high (i.e., above .95) in the large majority of conditions. This is not desirable and indicates that the standard errors are too large, resulting in a less than 5% Type I error rate. Thus the method is quite conservative. This was also found in Asparohouv and Muthén (2014) with the ML estimator. However, in some extreme conditions the coverage of factor means and factor variances was below the nominal rate of .95. This occurred for factor means when there were 3 groups, 43% noninvariance, and large differences in the thresholds. This also occurred for factor variances across all numbers of groups when there was large and medium noninvariance on 43% of the loadings.

Table 8.

Average Coverage of Loadings by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI

Amount and Magnitude of Noninvariance	**3g**	**9g**	**15g**	**Grand Total**
Location of Noninvariance: Loadings				
0.14	**0.969**	**0.970**	**0.970**	**0.970**
LG	0.968	0.969	0.970	0.970
MD	0.968	0.970	0.971	0.970
SM	0.970	0.971	0.970	0.970
0.29	**0.966**	**0.969**	**0.969**	**0.968**
LG	0.962	0.969	0.968	0.967
MD	0.966	0.969	0.968	0.968
SM	0.969	0.970	0.971	0.970
0.43	**0.965**	**0.969**	**0.969**	**0.969**
LG	0.959	0.968	0.968	0.967
MD	0.964	0.968	0.967	0.967
SM	0.971	0.971	0.971	0.971
Location of Noninvariance: Thresholds				
0.14	**0.970**	**0.971**	**0.971**	**0.971**
LG	0.970	0.970	0.970	0.970
MD	0.970	0.971	0.971	0.971
SM	0.970	0.970	0.971	0.971
0.29	**0.971**	**0.971**	**0.972**	**0.971**
LG	0.971	0.972	0.972	0.972
MD	0.971	0.971	0.971	0.971
SM	0.971	0.971	0.972	0.971
0.43	**0.971**	**0.971**	**0.971**	**0.971**
LG	0.971	0.971	0.971	0.971
MD	0.971	0.971	0.971	0.971
SM	0.970	0.970	0.972	0.971
Complete Invariance	**0.969**	**0.970**	**0.971**	**0.970**
Grand Total	**0.968**	**0.970**	**0.970**	**0.970**

Table 9.

Average Coverage of Thresholds by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI

Amount and Magnitude of Noninvariance	**3g**	**9g**	**15g**	**Grand Total**
Location of Noninvariance: Loadings				
0.14	**0.962**	**0.963**	**0.960**	**0.961**
LG	0.963	0.963	0.961	0.962
MD	0.962	0.963	0.961	0.961
SM	0.962	0.963	0.959	0.961
0.29	**0.963**	**0.963**	**0.961**	**0.962**
LG	0.963	0.963	0.961	0.962
MD	0.963	0.963	0.961	0.962
SM	0.962	0.963	0.960	0.961
0.43	**0.962**	**0.963**	**0.961**	**0.962**
LG	0.963	0.964	0.961	0.962
MD	0.963	0.963	0.961	0.962
SM	0.962	0.963	0.960	0.961
Location of Noninvariance: Thresholds				
0.14	**0.963**	**0.963**	**0.960**	**0.962**
LG	0.964	0.964	0.960	0.962
MD	0.963	0.963	0.961	0.962
SM	0.962	0.963	0.960	0.961
0.29	**0.959**	**0.957**	**0.954**	**0.955**
LG	0.960	0.958	0.954	0.956
MD	0.957	0.952	0.950	0.951
SM	0.960	0.960	0.957	0.959
0.43	**0.948**	**0.955**	**0.949**	**0.951**
LG	0.934	0.953	0.949	0.949
MD	0.949	0.950	0.944	0.946
SM	0.960	0.961	0.955	0.957
Complete Invariance	**0.961**	**0.962**	**0.959**	**0.961**
Grand Total	**0.960**	**0.961**	**0.958**	**0.961**

Table 10.

Average Coverage of Factor Means by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI*

Amount and Magnitude of Noninvariance	**3g**	**9g**	**15g**	**Grand Total**
Location of Noninvariance: Loadings				
0.14	**0.981**	**0.972**	**0.966**	**0.969**
LG	0.983	0.973	0.968	0.971
MD	0.981	0.971	0.966	0.969
SM	0.981	0.971	0.964	0.968
0.29	**0.980**	**0.975**	**0.967**	**0.971**
LG	0.981	0.979	0.970	0.974
MD	0.978	0.975	0.968	0.971
SM	0.982	0.972	0.965	0.968
0.43	**0.962**	**0.959**	**0.948**	**0.953**
LG	0.950	0.950	0.937	0.943
MD	0.961	0.959	0.947	0.952
SM	0.976	0.968	0.960	0.964
Location of Noninvariance: Thresholds				
0.14	**0.981**	**0.972**	**0.967**	**0.970**
LG	0.980	0.971	0.966	0.969
MD	0.981	0.972	0.967	0.970
SM	0.982	0.972	0.968	0.970
0.29	**0.971**	**0.963**	**0.956**	**0.960**
LG	0.972	0.962	0.955	0.959
MD	0.966	0.956	0.949	0.952
SM	0.976	0.970	0.965	0.967
0.43	**0.952**	**0.964**	**0.952**	**0.956**
LG	0.924	0.959	0.950	0.951
MD	0.956	0.957	0.943	0.949
SM	0.977	0.975	0.963	0.968
Complete Invariance	**0.982**	**0.971**	**0.964**	**0.968**
Grand Total	**0.972**	**0.967**	**0.960**	**0.963**

Note. Average coverage across all groups' factor means was calculated without the first group because the factor mean was set to zero in the first group across all conditions and not estimated in the model

Table 11.

Average Coverage of Factor Variances by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI*

Amount and Magnitude of Noninvariance	**3g**	**9g**	**15g**	**Grand Total**
Location of Noninvariance: Loadings				
0.14	**0.977**	**0.972**	**0.968**	**0.970**
LG	0.977	0.973	0.971	0.972
MD	0.977	0.972	0.967	0.970
SM	0.978	0.972	0.966	0.969
0.29	**0.977**	**0.975**	**0.971**	**0.973**
LG	0.977	0.977	0.976	0.976
MD	0.976	0.975	0.971	0.973
SM	0.978	0.972	0.968	0.970
0.43	**0.928**	**0.941**	**0.934**	**0.936**
LG	0.888	0.926	0.915	0.916
MD	0.924	0.937	0.928	0.931
SM	0.971	0.960	0.957	0.959
Location of Noninvariance: Thresholds				
0.14	**0.977**	**0.972**	**0.968**	**0.970**
LG	0.978	0.973	0.968	0.970
MD	0.979	0.973	0.969	0.971
SM	0.975	0.972	0.967	0.970
0.29	**0.978**	**0.972**	**0.970**	**0.971**
LG	0.979	0.973	0.970	0.972
MD	0.979	0.972	0.970	0.971
SM	0.977	0.972	0.969	0.971
0.43	**0.979**	**0.973**	**0.970**	**0.972**
LG	0.982	0.974	0.970	0.972
MD	0.981	0.973	0.969	0.971
SM	0.976	0.971	0.970	0.971
Complete Invariance	**0.980**	**0.973**	**0.967**	**0.970**
Grand Total	**0.970**	**0.968**	**0.964**	**0.966**

Note. Average coverage across all groups' factor variances was calculated without the first group because the factor variance was set to one in the first group across all conditions and not estimated in the model

Note. When calculating the average coverage across all groups' factor variances information from two conditions was not included because the estimated factor variance was too large to be printed: group 8 in the15g, 43%NI,LG magnitude, NI on the Loadings condition and group 12 in the 15g, 14%NI, LG, Thresholds condition

True and estimated factor means. Mplus provides the correlation between the true and estimated factor means. A correlation of one occurs when the rank order of the groups was completely recovered by the alignment. Though this does not provide any information about the bias in the estimation of the factor means, it does provide a weak test of factor mean recovery. Figure 4 shows the correlation between true and estimated factor means for each subscale by number of groups, percent of noninvariance, magnitude of noninvariance, and location of noninvariance. As can be seen on this figure, true and estimated factor mean correlations are always greater than .90, with slightly lower values as the number of groups increases. They are only lower than .98 when there is 29-43% noninvariance and the magnitude is large or medium. Further, the recovery of group rank order was more negatively impacted for the time management subscale than the intrinsic motivation subscale. Thus, it is expected that in these extreme conditions there is also bias and variability in the estimates, which I present next.

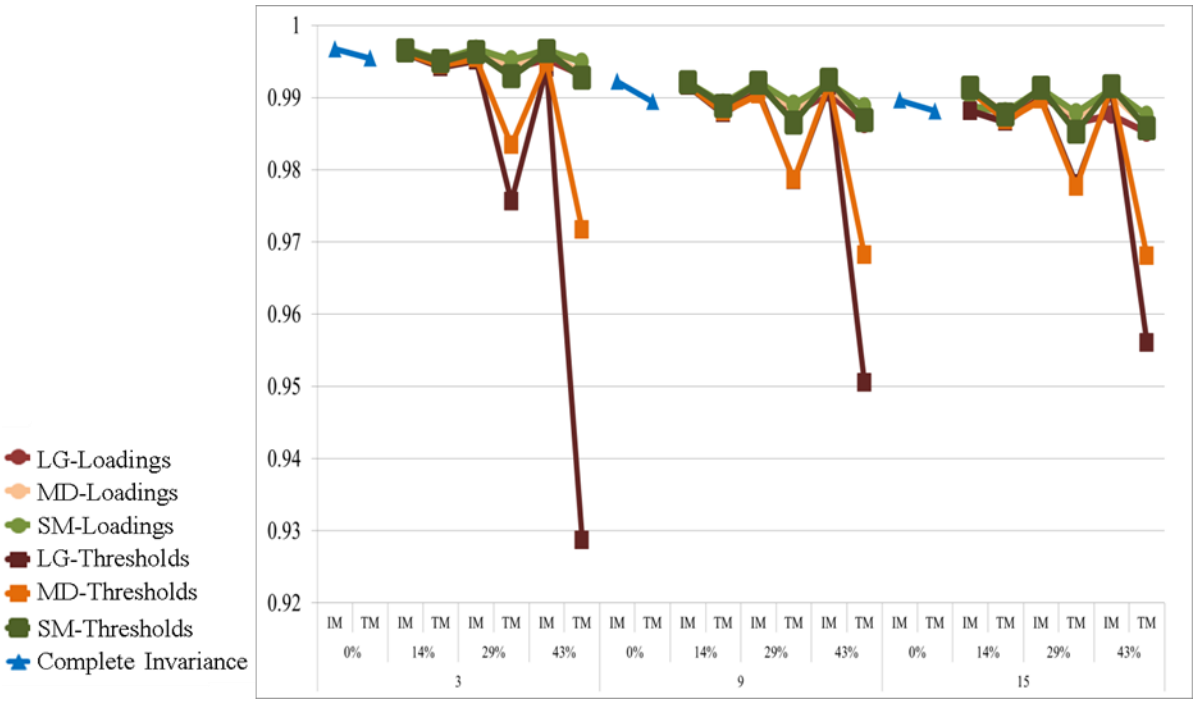

Figure 4. True and Estimated Factor Mean Correlations

Relative bias and M.S.E. summary. Tables 12-15 are similarly structured to the coverage tables and include the relative bias and M.S.E. of loadings, thresholds, factor means, and factor variances. In calculating the relative bias, the first group type was not included in the summary because the factor mean in group type 1 was set to zero. This created a situation where the denominator of the relative bias equation was zero and the solution was undefined. Further, some groups' information was excluded in calculating the average values for the factor variances, because the estimate of that group's factor variance was printed as "*********" in Mplus. When this occurred there is an explanatory note at the bottom of the table.

The relative bias and M.S.E. are small in most conditions for most parameters. For loadings, there was increased relative bias (greater than .05) when 43% of the loadings were noninvariant and the magnitude was large. This occurred across the 3, 9, and 15 groups conditions with a relative bias of approximately .06. Still, relative bias was never greater than .10 for the loadings. The results were similar for the thresholds, there was only increased relative bias when there was 43% noninvariance on the thresholds and the magnitude was large.

When considering the factor means and variances there was a greater effect of noninvariance with larger relative bias and MSE. Generally, when there was 43% noninvariance that was medium or large in magnitude, relative bias was increased dramatically, greater than .10. This was more severe when there was threshold noninvariance and a larger number of groups. There are a few instances where the relative bias is extreme, above 1.0. These values were primarily driven up by the incorrect estimation of a single group's factor mean in the large magnitude conditions. As an example, in the 15 groups, 14% large threshold noninvariance, one group (group 12) had an estimated factor mean of -77.5, which increased the average relative bias substantially. This type of extreme bias also occurred in a few other 15-group conditions,

details of which are included in the notes at the bottom of the table. This also occurred with the estimates of the factor variances. However, in these cases, the values were so large that they were not printed by Mplus. These instances are noted in the table.

Table 12.
Average Relative Bias and Mean Square Error of Loadings by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI

Amount and Magnitude of Noninvariance	3g		9g		15g		Grand Totals	
	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE
Location of Noninvariance: Loadings								
0.14	**0.011**	**0.024**	**0.010**	**0.023**	**0.008**	**0.025**	**0.009**	**0.024**
LG	0.010	0.025	0.009	0.024	0.007	0.023	0.008	0.023
MD	0.011	0.024	0.010	0.023	0.009	0.023	0.009	0.023
SM	0.011	0.023	0.010	0.023	0.009	0.029	0.010	0.026
0.29	**0.012**	**0.026**	**0.009**	**0.024**	**0.007**	**0.023**	**0.008**	**0.024**
LG	0.011	0.028	0.006	0.025	0.004	0.024	0.006	0.025
MD	0.012	0.025	0.010	0.024	0.008	0.023	0.009	0.023
SM	0.012	0.024	0.010	0.023	0.008	0.022	0.009	0.023
0.43	**0.049**	**0.034**	**0.044**	**0.030**	**0.042**	**0.029**	**0.043**	**0.030**
LG	0.069	0.043	0.062	0.036	0.060	0.034	0.062	0.036
MD	0.051	0.033	0.046	0.030	0.045	0.029	0.046	0.029
SM	0.026	0.025	0.023	0.025	0.021	0.024	0.022	0.024
Location of Noninvariance: Thresholds								
0.14	**0.010**	**0.024**	**0.010**	**0.024**	**0.007**	**0.023**	**0.008**	**0.024**
LG	0.010	0.025	0.010	0.024	0.007	0.025	0.008	0.025
MD	0.010	0.024	0.010	0.024	0.006	0.023	0.008	0.023
SM	0.010	0.024	0.009	0.023	0.006	0.023	0.008	0.023
0.29	**0.011**	**0.024**	**0.009**	**0.024**	**0.007**	**0.023**	**0.008**	**0.024**
LG	0.010	0.025	0.010	0.025	0.007	0.024	0.008	0.024
MD	0.011	0.024	0.010	0.024	0.007	0.023	0.008	0.023
SM	0.010	0.024	0.009	0.023	0.007	0.023	0.008	0.023
0.43	**0.010**	**0.024**	**0.009**	**0.024**	**0.007**	**0.023**	**0.008**	**0.024**
LG	0.009	0.025	0.010	0.025	0.006	0.024	0.008	0.024
MD	0.011	0.024	0.009	0.024	0.007	0.023	0.008	0.024
SM	0.010	0.024	0.008	0.023	0.007	0.023	0.008	0.023
Complete Invariance	**0.010**	**0.023**	**0.009**	**0.023**	**0.008**	**0.025**	**0.008**	**0.024**
Grand Totals	**0.017**	**0.026**	**0.015**	**0.025**	**0.013**	**0.024**	**0.014**	**0.025**

Table 13.

Average Relative Bias and Mean Square Error of Thresholds by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI

Amount and Magnitude of Noninvariance	3g		9g		15g		Grand Totals	
	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE
Location of Noninvariance: Loadings								
0.14	**0.011**	**0.039**	**0.011**	**0.035**	**0.012**	**0.034**	**0.012**	**0.035**
LG	0.011	0.038	0.011	0.035	0.012	0.033	0.012	0.034
MD	0.011	0.039	0.011	0.035	0.012	0.034	0.012	0.035
SM	0.011	0.039	0.011	0.035	0.012	0.036	0.012	0.036
0.29	**0.011**	**0.039**	**0.012**	**0.035**	**0.012**	**0.033**	**0.012**	**0.034**
LG	0.011	0.038	0.012	0.034	0.012	0.033	0.012	0.034
MD	0.011	0.038	0.012	0.035	0.012	0.033	0.012	0.034
SM	0.011	0.039	0.011	0.035	0.012	0.034	0.012	0.035
0.43	**0.011**	**0.039**	**0.011**	**0.035**	**0.012**	**0.034**	**0.011**	**0.035**
LG	0.010	0.039	0.011	0.034	0.011	0.035	0.011	0.035
MD	0.011	0.039	0.011	0.035	0.012	0.033	0.012	0.034
SM	0.011	0.039	0.011	0.035	0.012	0.034	0.012	0.035
Location of Noninvariance: Thresholds								
0.14	**0.009**	**0.046**	**0.010**	**0.047**	**0.010**	**0.051**	**0.010**	**0.049**
LG	0.007	0.055	0.007	0.060	0.007	0.072	0.007	0.066
MD	0.009	0.042	0.011	0.045	0.012	0.043	0.012	0.044
SM	0.010	0.040	0.010	0.036	0.011	0.037	0.010	0.037
0.29	**0.021**	**0.053**	**0.017**	**0.051**	**0.014**	**0.053**	**0.016**	**0.052**
LG	0.012	0.066	0.008	0.066	0.003	0.073	0.006	0.070
MD	0.029	0.050	0.025	0.050	0.019	0.048	0.022	0.049
SM	0.021	0.042	0.019	0.037	0.019	0.038	0.019	0.038
0.43	**0.024**	**0.064**	**0.018**	**0.061**	**0.015**	**0.067**	**0.017**	**0.064**
LG	0.077	0.091	0.067	0.092	0.060	0.107	0.064	0.100
MD	-0.015	0.058	-0.019	0.054	-0.023	0.054	-0.021	0.054
SM	0.009	0.042	0.008	0.037	0.008	0.038	0.008	0.038
Complete Invariance	**0.010**	**0.039**	**0.011**	**0.035**	**0.012**	**0.037**	**0.011**	**0.037**
Grand Totals	**0.014**	**0.046**	**0.013**	**0.044**	**0.012**	**0.045**	**0.013**	**0.045**

Table 14.

Average Relative Bias and Mean Square Error of Factor Means by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI*

Amount and Magnitude of Noninvariance	3g		9g		15g		Grand Totals	
	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE
Location of Noninvariance: Loadings								
0.14	**0.000**	**0.012**	**-0.013**	**0.010**	**-0.018**	**0.010**	**-0.014**	**0.010**
LG	0.003	0.013	-0.012	0.010	-0.016	0.010	-0.013	0.010
MD	0.000	0.012	-0.014	0.010	-0.019	0.010	-0.015	0.010
SM	-0.002	0.011	-0.014	0.010	-0.019	0.010	-0.016	0.010
0.29	**0.002**	**0.014**	**-0.012**	**0.010**	**-0.015**	**0.010**	**-0.012**	**0.011**
LG	0.008	0.016	-0.007	0.011	-0.007	0.011	-0.005	0.012
MD	0.000	0.014	-0.014	0.010	-0.018	0.010	-0.015	0.011
SM	-0.003	0.011	-0.015	0.010	-0.019	0.010	-0.016	0.010
0.43	**-0.048**	**0.016**	**-0.061**	**0.012**	**-0.531**	**0.013**	**-0.321**	**0.013**
LG	-0.067	0.020	-0.082	0.015	-1.485[a]	0.016	-0.860	0.016
MD	-0.053	0.015	-0.065	0.012	-0.070	0.012	-0.066	0.012
SM	-0.023	0.012	-0.034	0.010	-0.038	0.010	-0.035	0.010
Location of Noninvariance: Thresholds								
0.14	**0.000**	**0.013**	**-0.013**	**0.011**	**-1.323**	**0.011**	**-0.739**	**0.011**
LG	-0.007	0.013	-0.022	0.011	-3.949[a]	0.012	-2.202	0.012
MD	0.001	0.013	-0.013	0.011	-0.013	0.010	-0.012	0.011
SM	0.007	0.012	-0.005	0.010	-0.006	0.010	-0.004	0.010
0.29	**-0.022**	**0.020**	**-0.027**	**0.015**	**-0.010**	**0.014**	**-0.017**	**0.015**
LG	-0.032	0.025	-0.035	0.017	-0.023	0.015	-0.028	0.017
MD	-0.027	0.022	-0.032	0.017	-0.017	0.016	-0.023	0.017
SM	-0.007	0.013	-0.013	0.011	0.011	0.011	0.001	0.011
0.43	**0.130**	**0.034**	**0.116**	**0.022**	**0.123**	**0.020**	**0.122**	**0.022**
LG	0.194	0.058	0.158	0.031	0.152	0.027	0.159	0.032
MD	0.140	0.032	0.142	0.022	0.145	0.021	0.144	0.023
SM	0.055	0.014	0.049	0.012	0.073	0.012	0.063	0.012
Complete Invariance	**0.000**	**0.011**	**-0.011**	**0.010**	**-0.455[b]**	**0.011**	**-0.256**	**0.011**
Grand Totals	**0.010**	**0.018**	**-0.002**	**0.013**	**-0.304**	**0.013**	**-0.168**	**0.013**

Note. The first group type, with a mean 0 is not included in the calculation of the average relative bias and MSE.

[a] These estimates were driven up by a single group's estimated factor mean, group 12.

[b] This estimate was driven up by group 8, which had an extreme estimate

Table 15.
Average Relative Bias and Mean Square Error of Factor Variances by Number of Groups, Location of NI, Amount of NI, and Magnitude of NI*

Amount and Magnitude of Noninvariance	3g		9g		15g		Grand Totals	
	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE	Relative Bias	MSE
Location of Noninvariance: Loadings								
0.14	**0.015**	**0.065**	**0.014**	**0.052**	**0.017**	**0.048**	**0.016**	**0.050**
LG	0.020	0.074	0.018	0.057	0.020	0.051	0.019	0.055
MD	0.014	0.064	0.013	0.051	0.015	0.047	0.014	0.049
SM	0.010	0.057	0.011	0.047	0.015	0.046	0.013	0.047
0.29	**0.022**	**0.094**	**0.020**	**0.062**	**0.023**	**0.056**	**0.022**	**0.061**
LG	0.038	0.133	0.031	0.078	0.034	0.070	0.033	0.078
MD	0.019	0.088	0.017	0.059	0.018	0.054	0.018	0.058
SM	0.009	0.060	0.012	0.048	0.016	0.045	0.014	0.048
0.43	**-0.072**	**0.103**	**-0.051**	**0.064**	**-0.046**	**0.057**	**-0.050**	**0.064**
LG	-0.103	0.155	-0.079	0.085	-0.075	0.075	-0.079	0.085
MD	-0.083	0.096	-0.058	0.062	-0.053	0.055	-0.057	0.061
SM	-0.031	0.059	-0.017	0.046	-0.011	0.043	-0.015	0.045
Location of Noninvariance: Thresholds								
0.14	**0.013**	**0.056**	**0.012**	**0.047**	**0.018**	**0.044**	**0.016**	**0.046**
LG	0.013	0.057	0.012	0.048	0.018	0.046	0.016	0.047
MD	0.013	0.057	0.012	0.047	0.018	0.044	0.016	0.046
SM	0.013	0.056	0.013	0.047	0.018	0.043	0.016	0.045
0.29	**0.012**	**0.056**	**0.013**	**0.047**	**0.018**	**0.044**	**0.016**	**0.046**
LG	0.014	0.057	0.013	0.048	0.019	0.045	0.016	0.047
MD	0.011	0.056	0.013	0.047	0.017	0.044	0.015	0.046
SM	0.011	0.054	0.014	0.047	0.016	0.043	0.015	0.045
0.43	**0.014**	**0.057**	**0.014**	**0.048**	**0.018**	**0.044**	**0.016**	**0.047**
LG	0.017	0.060	0.013	0.049	0.020	0.046	0.017	0.048
MD	0.012	0.056	0.014	0.047	0.018	0.045	0.016	0.046
SM	0.012	0.055	0.015	0.047	0.017	0.043	0.016	0.046
Complete Invariance	**0.013**	**0.055**	**0.014**	**0.046**	**0.017**	**0.044**	**0.015**	**0.045**
Grand Totals	**0.001**	**0.071**	**0.004**	**0.053**	**0.008**	**0.049**	**0.006**	**0.052**

Note. The first group's factor variance was fixed to one in each condition, that group is not included in the calculations. Information from the 15g set was excluded due to extreme values: group 8 in the 43% NI, LG, Loadings, group 12 in the 14% NI, LG, Thresholds, group 13 in the 14% NI, SM, Loadings, and group 8 from the 0% NI condition.

Relative bias and M.S.E. detailed. The summary tables provide a way to identify what conditions are the most problematic. But, because these tables average across groups, items, and

both factors in the model, it is not clear if a certain group, item, or factor is more problematic. In the following section, I report a more detailed analysis for the extreme conditions. This provided an avenue for identifying potential causes of estimation issues across the conditions.

Figure 5 shows the MSE and relative bias for the loadings and thresholds for the first three groups for a few select items. These graphs only include output from the 29% and 43% noninvariance conditions, where the magnitude was large, and the noninvariance was located on those parameters, as these were the conditions that showed increased bias from the summary tables. For example, in the following figure showing the MSE and relative bias in the loadings, the bars represent the average MSE and relative bias across all the estimates from 29 and 43% large loading noninvariance for the first three groups for three items: Y10, Y3, and Y5. Of these three items, Y10 and Y3 are noninvariant in group two, Y5, however, in the 43% noninvariance condition is completely noninvariant (i.e., different across all three group types), this is marked with a star in the figure. The MSE and relative bias for all of the thresholds and loadings from these extreme conditions are included in Appendix C.

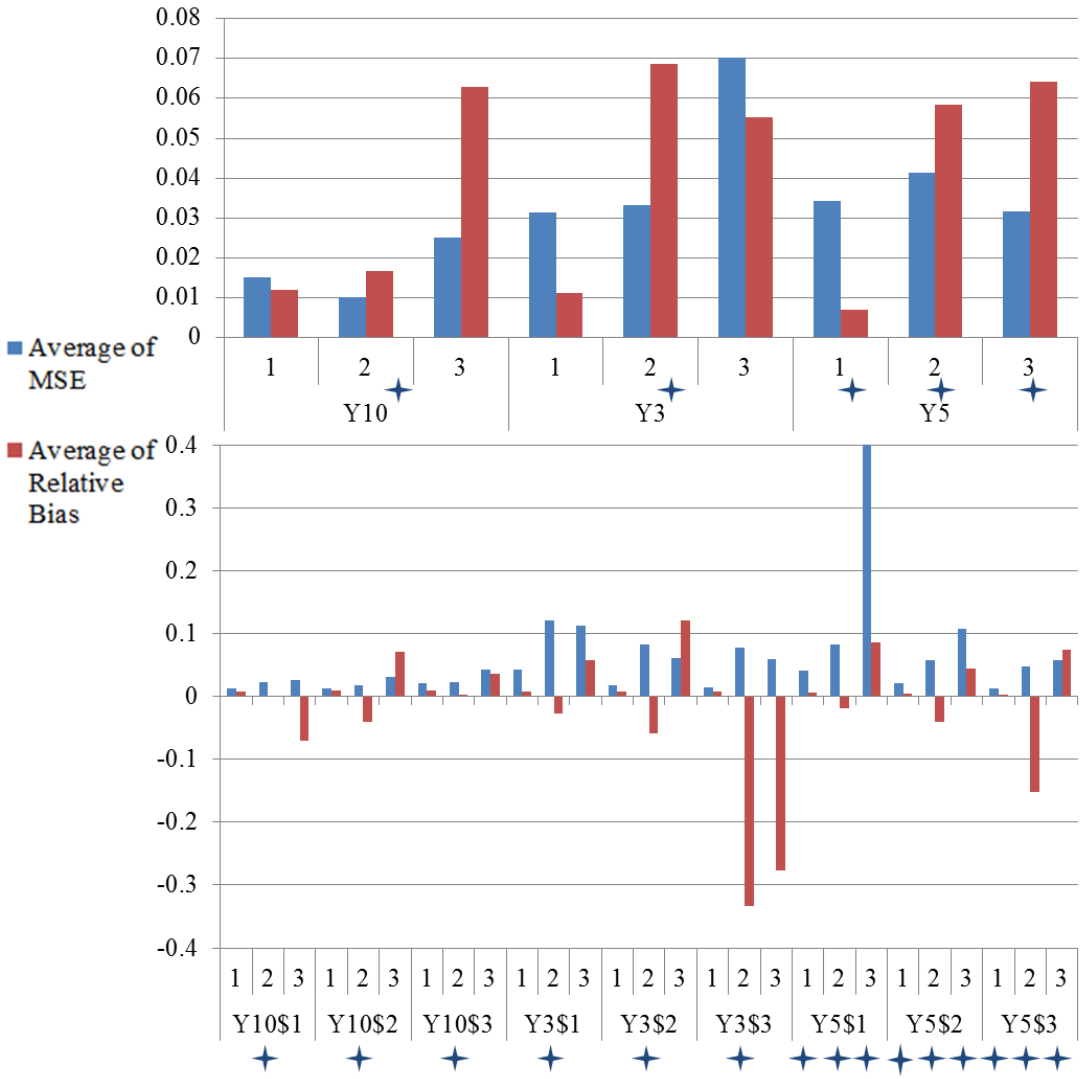

Figure 5. MSE and Relative Bias of Select Items. MSE for Y5$1 in group 3 was 4.0, but the scale of the graph was restricted so the differences could be seen across the other parameters.

The detailed item parameter figures showed that some parameters in some groups were better estimated than others. Item 10, for example, has consistently less bias and MSE than the other items for the loadings and the thresholds. Relative bias and MSE were quite low on average; however in these extreme conditions, some individual parameters were severely misestimated.

Figure 6 shows the MSE (top) and relative bias (bottom) of the factor means in more detail for the conditions where there were 29% and 43% noninvariant items of a medium or large magnitude on the loadings or thresholds. This figure shows that the average of the MSE and relative bias were driven up considerably by the bias in the estimate of the time management factor mean for group 2. The primary difference between groups 2 and 3 is the factor means and variances (group type 2, $\alpha = .3$, $\psi = 1.5$, group type 3, $\alpha = 1$, $\psi = 1.2$). Again, the bias and MSE are higher when the noninvariance was located on the thresholds, with only negligible values of relative bias and MSE when the noninvariance was on the loadings.

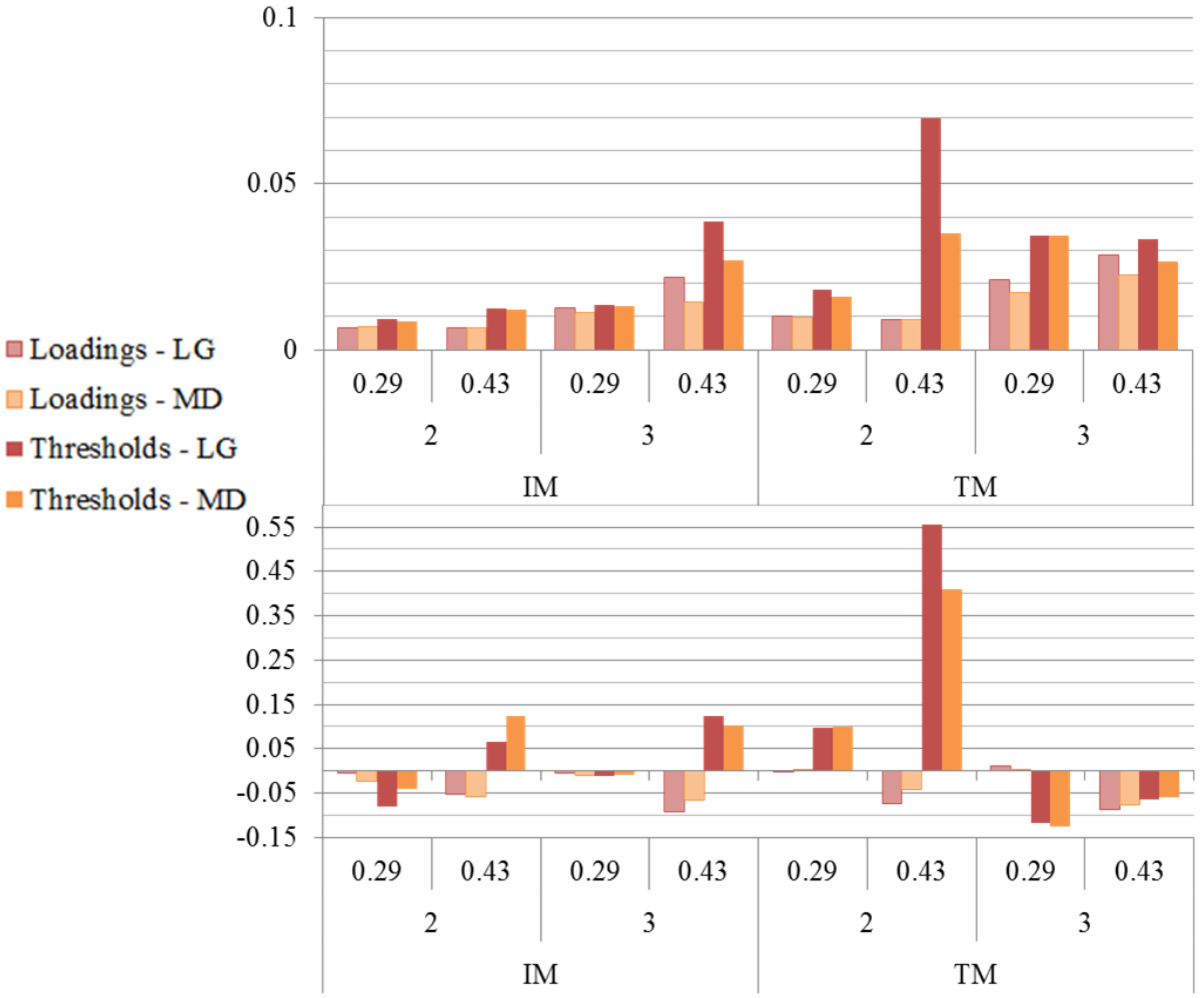

Figure 6. MSE and Relative Bias of Factor Means in Select Conditions.

Noninvariance testing procedure. First I considered what percentage of the time, across replications, the procedure listed groups as invariant when they were simulated as invariant. To create a summary of this I considered items Y7 and Y14, which had invariant thresholds and loadings across all conditions for group types 1 and 2 (see table x for pattern of noninvariance). Groups 1 and 2 were found invariant across replications for items Y7 and Y14 at least 99% of the time, regardless of which condition or parameter (loading or threshold). Thus, the procedure does not have a substanial error rate when detecting invariance of invariant parameters. This error rate is not surprising given the stringent .001 alpha value used in the algorythm. In considering the accurate detection of noninvariant parameters, I plotted the hit rate of the testing procedure in Figure 7 for a select group of items. These graphs show the percentage of times the loadings (top) and thresholds (bottom) were identifed as noninvariant by the procedure. The x-axis shows the parameter and then group type. The bars represent the magnitude and percentage of noninvariance. The parameters with the blue star indicate that the parameter was simulated as noninvariant in that group, across all conditions.

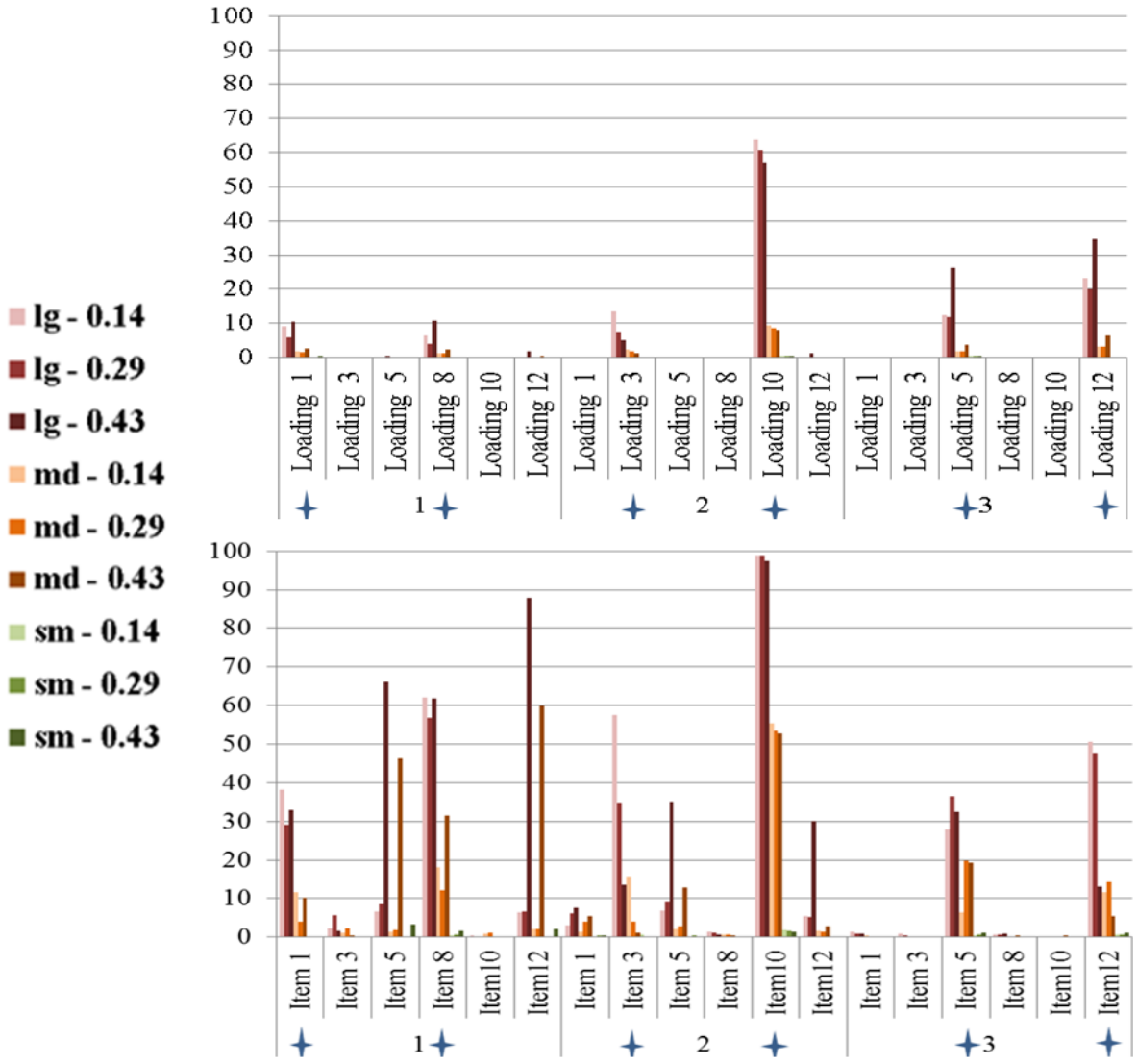

Figure 7. Percentage of Replications Parameter was Flagged as Noninvariant.

As can be seen in the top graph in Figure 7, which displays the percentage of times noninvariant loadings were declared noninvariant by the procedure, the testing is extremely conservative. Even with larger magnitude of differences, the loading was deemed as invariant. I averaged these results across the number of groups conditions because the results were similar if there were 3, 9, or 15 groups. The procedure seemed to perform the best with item 10, flagging the large noninvariant loadings difference approximately 65% of the time. The lower graph in Figure 7 shows the results averaged across the three thresholds and number of groups. The results for the thresholds are more favorable, with the procedure flagging the noninvariant thresholds in the large conditions a majority of the time. Of the items plotted in these graphs, the procedure performs best with item 10. In general, the procedure performs better for some items than for others; appendix D includes the hit rate for all thresholds by group type for all items in the 29 and 43% large magnitude threshold noninvariance conditions.

Another indicator of noninvariance that the procedure provides is the R^2 for each item. A completely invariant item should have an R^2 value close to 1, which means that all the variance in the item parameter is explained by underlying group mean differences. Thus, once the item parameters have been aligned, the groups will be comparable. A noninvariant item should have a value close to zero. Figure 8 displays the R^2 values for a subset of items in the 3 groups, 23 and 43% noninvariant thresholds, when the magnitude was medium or large. The R^2 across the conditions for all three thresholds is averaged in the figure, representing an overall R^2 for the thresholds for that item. Items 7 and 14 are indicated on the graph because they are completely invariant in all of the 29% conditions, but not the 43% conditions. Thus, in the 29% conditions we would expect and R^2 to be very close to one, but lower in the 43% conditions. The rest of the items shown are noninvariant in at least one group in all of the conditions.

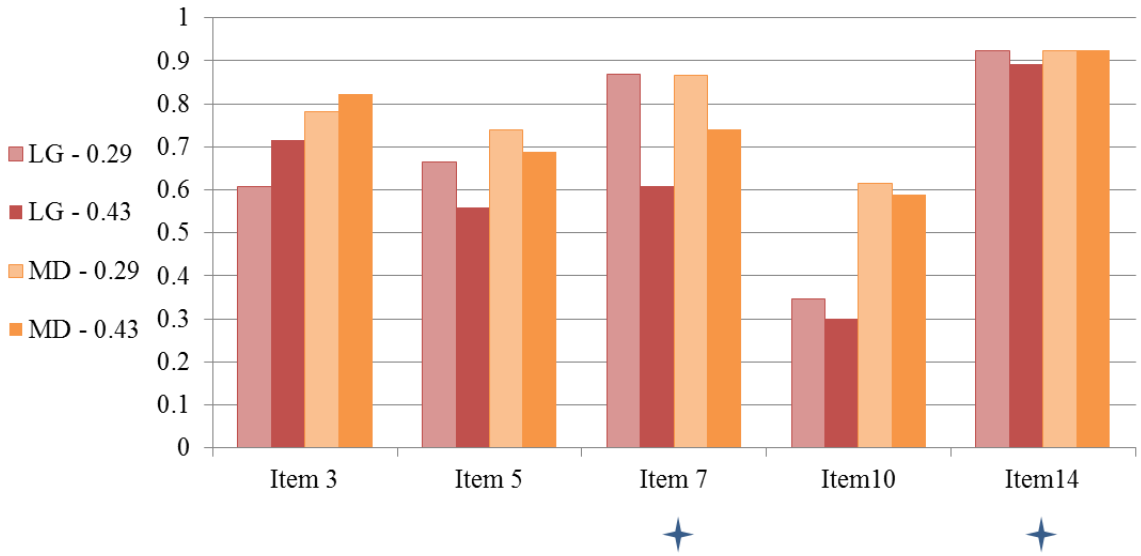

Figure 8. R^2 Values for a Subset of Items

The graph shows that the most severely noninvariant item, 5, does have lower R^2 values, however, item 10, has the lowest. The R^2 did show some sensitivity to item 7 being noninvariant in the 43% conditions, but not the 29%, as the R^2 is lower in the 43% conditions. However, this same pattern is not reflected with item 14. As with other estimates in the model, such as the factor means and noninvariance detection, the accuracy of the R^2 varies quite a bit depending on the item.

Investigating item characteristic differences. The summary tables of relative bias and MSE give some indication that, overall, the loadings and thresholds are generally estimated with little bias and variability across replications. However, in looking at individual items, it is apparent that there is variability in the MSE, relative bias, noninvariance testing performance, and R^2 across different parameters. Item 10, for example, clearly has the lowest bias and MSE, with the MSE being much larger for the other items, double if not triple. So, even though, overall, the bias and MSE are low, for individual parameters, some of the estimates fluctuate, which could be causing problems with the testing procedure. Another consideration is the

differences in item characteristics. Figure 9 and Figure 10 show the item characteristic curves for each item. As can be seen, some items cover the range of the latent factor better than others, and those items also seem to be better estimated and flagged as noninvariant by the testing procedure. Perhaps the difference in the R^2 values for items 7 and 14 is related to the mean structure of the items. Another example of this is item 10, which is one of the least extreme in its mean structure; an artifact of the general tendency of people to rate noncognitive scales favorably. Relatedly, in the real data, there were more responses in the first and second category for item 10 in comparison to other items, thus more information to estimate the thresholds parameters. For comparison purposes, the starting values for the thresholds of these items are also shown in Table 16.

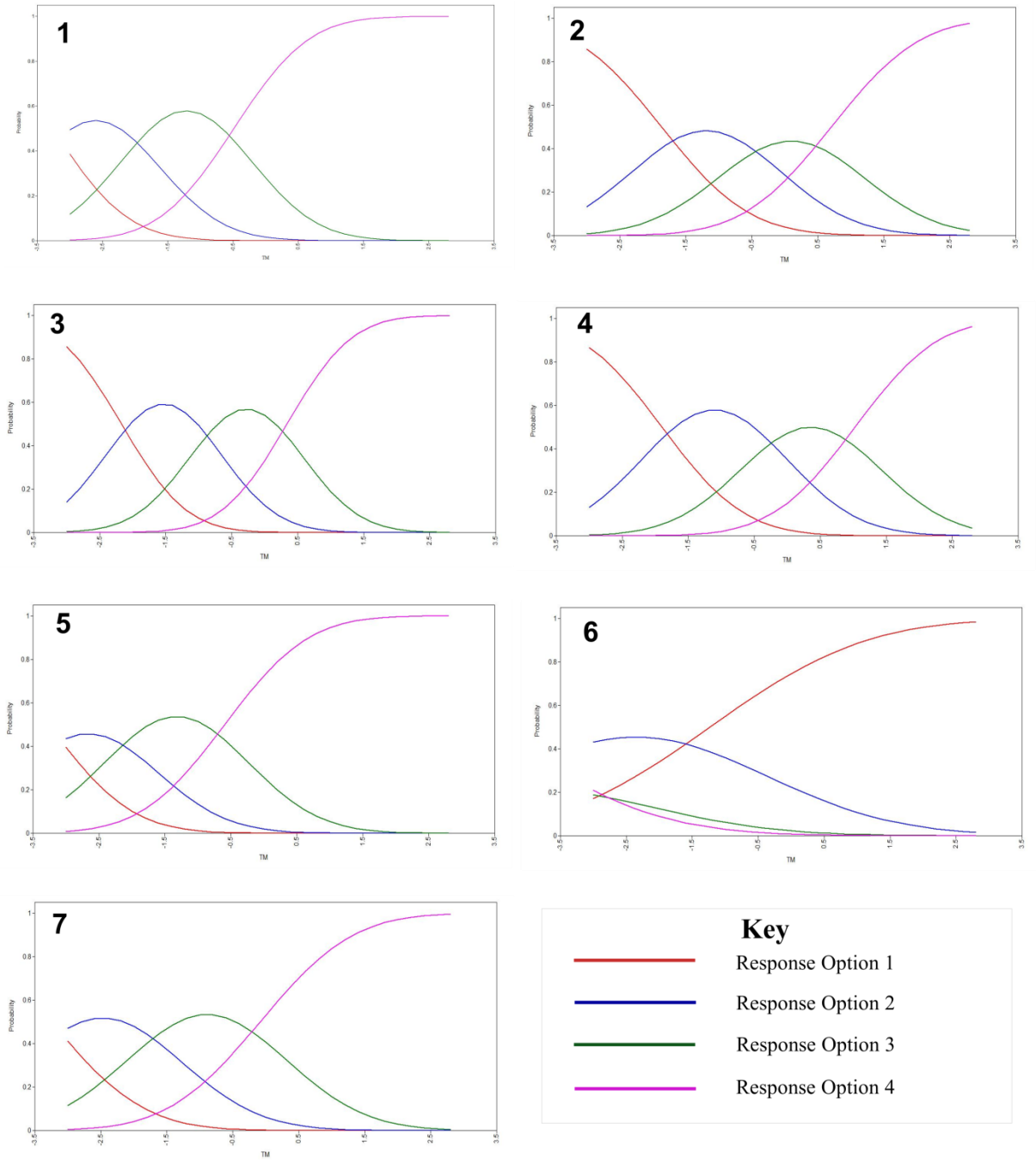

Figure 9. Item Characteristic Curves for TM.

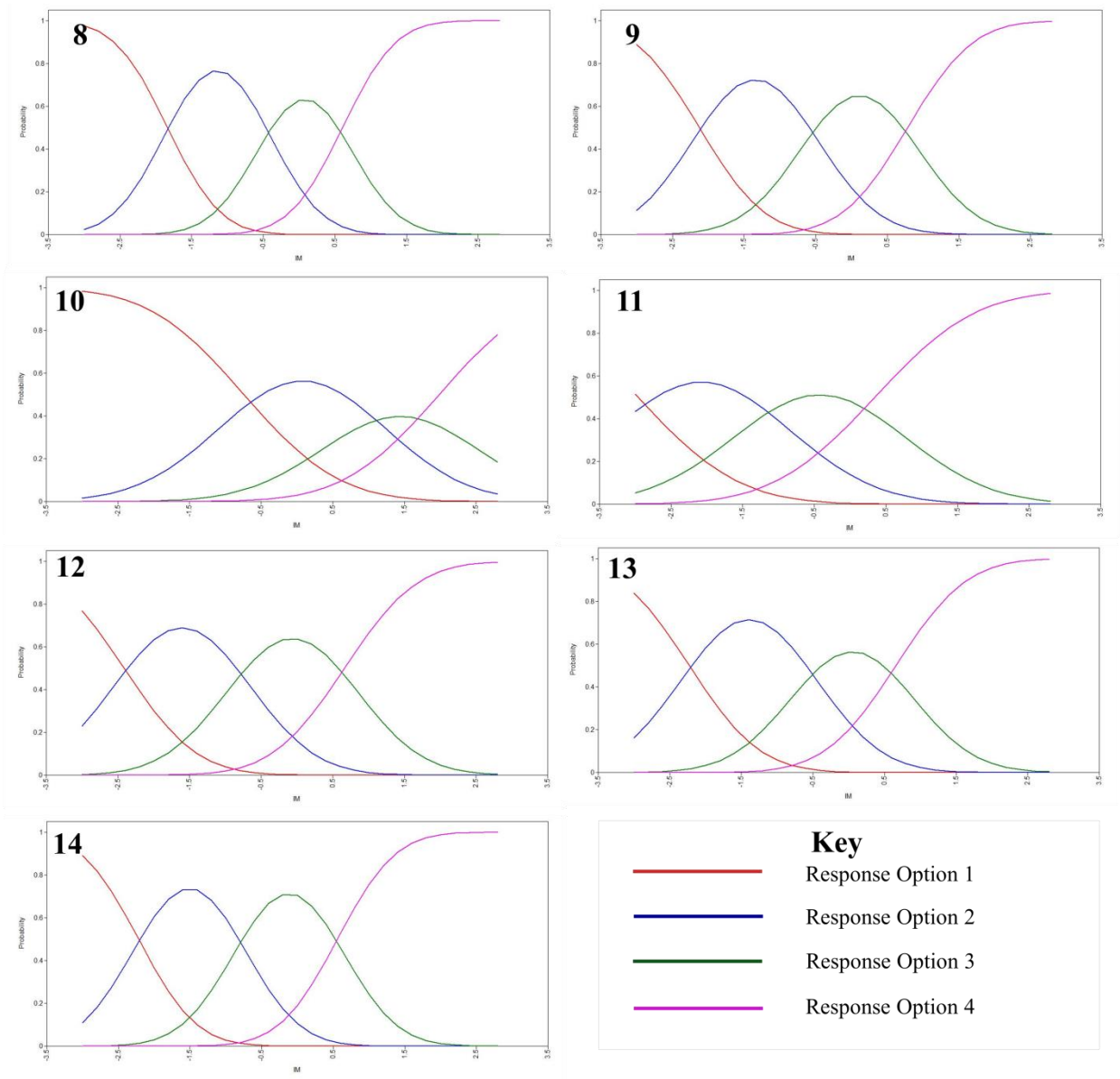

Figure 10. Item Characteristic Curves for IM.

Table 16.
Starting Values for Items 3, 5 and 10

Subscale	Parameter	Value
TM	Y3$1	-2.796
TM	Y3$2	-1.141
TM	Y3$3	0.432
TM	Y5$1	-3.284
TM	Y5$2	-2.066
TM	Y5$3	-0.597
TM	Y7$1	-2.939
IM	Y10$1	-0.706
IM	Y10$2	0.85
IM	Y10$3	1.892

Further investigation of the testing procedure. In considering the performance of the testing procedure numerous questions arise. Is the lack of detection due to bias, such that the parameters are misestimated and the differences are not uncovered? Or, are the parameters appropriately estimated, but the procedure has such a stringent alpha level, such that even larger differences are not found significant? There are two points in the algorithm where an error could occur: the pairwise comparisons used to create the invariance set of groups for a given parameter and the comparison of individual parameters to the average of the invariance set. Again, I took a closer look at a few items from the 3 groups, 43% nonivariant thresholds condition, where the magnitude of the differences was large. I chose this condition to investigate more thoroughly because the procedure performed favorably for item 10, but not for the other items. Table 17 shows the simulated estimated and simulated invariance set, the simple average of all three groups, the average of the pairwise comparison differences between noninvariant parameters, their average p-value, and the R^2 value for the same three items discussed above.

Table 17.
Estimated and True Invariant Sets Averages and Pairwise Difference Averages

Parameter		Estimated Invariant Average	Simulated Invariant Set Average	Simple Average of all Groups	Estimated Average Noninvariant Difference [a]	Pairwise Difference p-value	R^2
Item 3	Threshold 1	-3.046	-2.796	-3.063	0.521	0.379	0.690
	Threshold 2	-1.358	-1.141	-1.407	0.518	0.295	0.723
	Threshold 3	0.252	0.432	0.165	0.525	0.287	0.731
Item 5	Threshold 1	-3.377	NA	-3.284	1.201	0.140	0.551
	Threshold 2	-1.955	NA	-2.066	1.129	0.074	0.560
	Threshold 3	-0.113	NA	-0.060	1.117	0.056	0.564
Item 10	Threshold 1	0.685	-0.706	-0.973	0.832	0.011	0.316
	Threshold 2	-0.894	0.850	0.583	0.845	0.011	0.294
	Threshold 3	-1.924	1.892	1.625	0.848	0.018	0.292

[a] These averages are for conditions where the simulated difference was .80

Table 17 shows that the issue lies in the creation of the invariance set. Again, there is a contrast between item 10 and the other items. It is only for item 10 that the invariant set is closer to the simulated values, whereas for items 1 and 5, it is closer to the simple average of all three groups, even though there are noninvariant groups that should have been rejected from the invariance set. In the case of item 3, the pairwise differences were underestimated (less than .80, which was the true value), whereas for item 5 they are overestimated. However, for both items, the pairwise differences are non-significant. Only item 10 has accurate pairwise differences and p-values. Further, it does seem that the R^2 value is sensitive to noninvariance, but like the other indicators of noninvariance, sensitive to the misestimation that occurs in extreme conditions. In the table it would be desirable if all of the R^2 values were low, with 5 being the lowest, as it is noninvariant across all three groups.

From Table 17, it seems that the cause of the inability for the testing procedure to detect noninvariance was a result of the creation of the invariance set of groups. The inclusion of noninvariant groups into the invariance set seems to be from both the misestimation of group differences and power to detect differences between groups. To further explore the issue of

power, I plotted the estimated pairwise differences and their associated p-value for the same three items, 3, 5 and 10, for the second thresholds, from the 3-group, 43% large threshold noninvariance condition (Figure 11). The first row of plots shows the full range of differences and their associated p-values, while the second row shows a zoomed-in plot of differences for which the p-values were under the typical alpha of .05. These plots show the p-value that resulted from the pairwise test of differences between item parameters. If this p-value was less than .001, the group's item parameter was calculated as a part of the invariance set.

In line with the other results, there is more dispersion of estimated group differences for items 3 and 5, compared to item 10. Again, item 10 had less bias and variability in estimation. The plots also show that even large differences in thresholds, greater than .80, were routinely found nonsignificant, per the .001 criterion. From these graphs it seems that the group differences need to be greater than 1.0 to routinely be significant at the .01 level, with even more extreme differences at the .001. A less conservative alpha value may be necessary to flag substantial differences in item parameters. Based on these graphs, even an alpha value of .05, which does not correct for the number of comparisons, would increase the accuracy of the procedure without flagging trivial amounts of noninvariance.

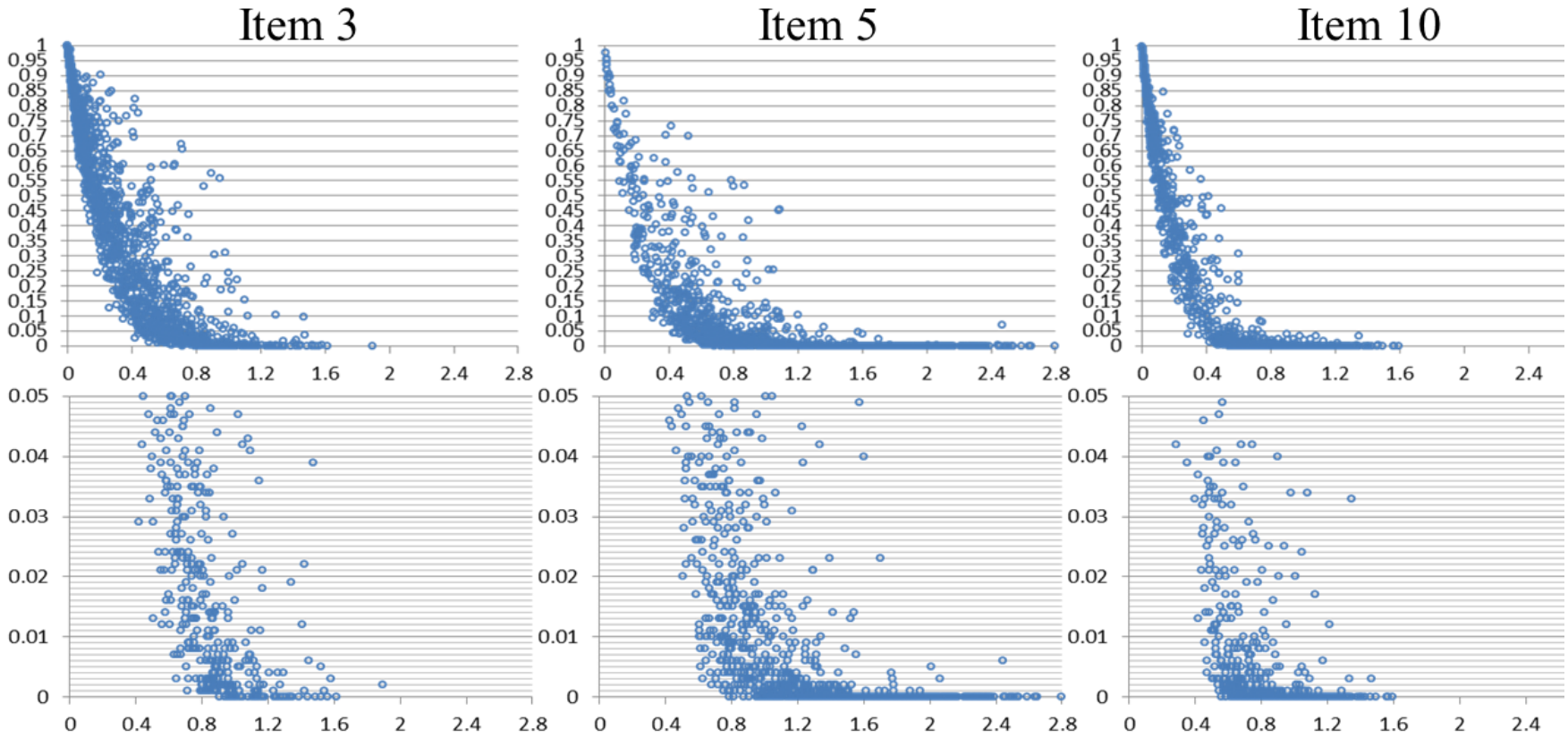

Figure 11. P-values by Pairwise Differences for Select Conditions.

Chapter 5: Discussion

The alignment method is a new method that has the potential to solve a complicated applied problem: comparing many groups on scales that have polytomous items. When comparing groups one must assume that the measurement of the construct is stable across groups so that differences in measurement are not confounded with differences in the construct. Traditional measurement invariance testing (i.e., DIF detection) methods can accommodate a few groups with these types of items, but become problematic when there are many groups. The alignment method addresses these problems by estimating the multiple-group factor models simultaneously and testing for noninvariance in an automated fashion. Thus, understanding how well the method performs under various conditions is critical for implementing it in practice.

The alignment is a very new method, with a new set of assumptions to understand and test. Of primary concern is the assumption of minimal measurement noninvariance. Asparohouv and Muthén (2013a) state that the component loss function used in the estimation of the group specific measurement models works well when there are many approximately invariant parameters and a few substantially noninvariant parameters. Naturally this leads the researcher to ask, "How much noninvariance is too much?"

The question of how much noninvariance is too much is not a new one, as many have posed these questions in regards to more traditional measurement invariance testing methods. Generally this question remains unanswered, and even more so for the alignment, which uses a different function, base model, and methodology completely. Muthén and Asparohouv (2014) have suggested that less than 25% of items should be noninvariant for the alignment to yield trustworthy results. This study was designed in particular to further explore that type of recommendation and investigate how extreme the conditions must be before problems arise. I

varied the number of noninvariant items, the magnitude of the differences between groups, and the location of the noninvariance. Through this combination of factors I provide initial recommendations about how much noninvariance is too much when items are polytomous.

Measurement Model Estimates

Overall the method was excellent at recovering the true parameters and produced estimates with little bias. Figure 12 shows the conditions of the study and an X indicates conditions where substantial decreases in coverage were observed, as well as increased bias. In this study, the method uniformly performed well when the magnitude of noninvariance was small, even if 43% of the items were noninvariant. So in considering the assumption of a majority of approximately invariant parameters, it seems that the differences simulated in the small conditions (loadings = -.10, thresholds = -.20) possess that property. Across all of the different estimates there were generally only substantial issues in conditions with 29-43% noninvariance of a medium or large magnitude, and this occurred if the noninvariance was on the loadings or the thresholds. Thus, it seems reasonable that the recommendation made by Muthén and Asparohouv (2014) suggesting fewer than 25% noninvariant items can be extended for polytomous data. However, it is important to note that this is in regards to the measurement model estimates, not the testing procedure. So, if researchers know the true number of noninvariant parameters, and those are less than 25% of items, the estimated loadings, thresholds, factor means, and factor variances are typically recovered well and with little bias.

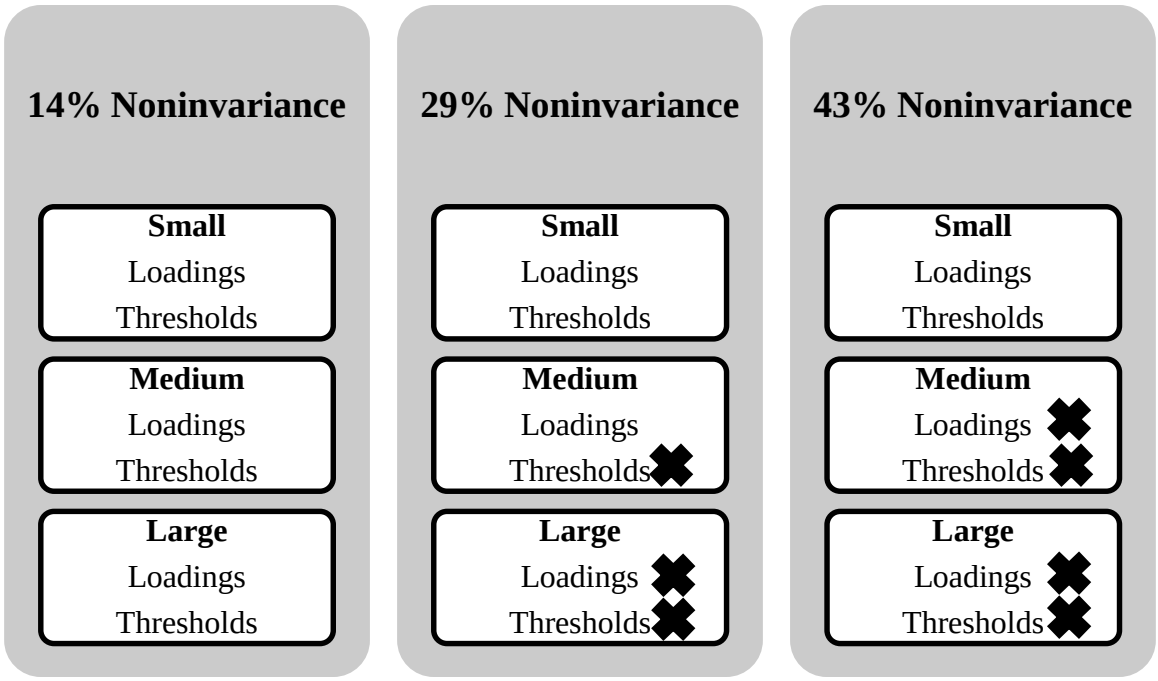

Figure 12. Overall Summary of Conditions.

The items experienced low coverage and increased bias when the noninvariance was on that parameter type. For example, when there was loading noninvariance, the estimation of the loadings exhibited increased bias and MSE. However, even in extreme cases of loading and threshold noninvariance the relative bias in associated parameters was relatively small, less than .10. Increases in noninvariance did cause higher bias and variability in the estimates, though it appears minor when averaged across all items. When looking at a few items and groups in depth, I observed variability in estimation across items and groups, observing that some items are better estimated than others and in certain groups. Through an investigation of the ICCs for each item, it seems that items with more coverage across the latent factor are better estimated. As in other methods, extremely negative thresholds may cause problems for the alignment. Item estimation issues seem to carry over into the testing procedure, causing a lack of power to detect substantial noninvariance.

Another notable finding about the item parameters was that groups with larger factor variances exhibited higher MSE in measurement model estimates. In looking through all item bias and MSE higher values were routinely observed in the 3rd group, which has the largest factor variance. The simulated factor means and variances were the same as Asparohouv and Muthén (2014). However, in their paper they only report estimates from group 2, which had a lower factor variance. Thus, it is not clear if they saw the same effect of group factor variance on item parameters.

In regards to factor means and variances, threshold noninvariance had the most substantial impact, which is consistent with other findings (Steinmetz, 2013). Medium and large threshold noninvariance on 29-43% of the items caused decreased true and estimated factor mean correlations as well as increased bias in factor means and variances. Occasionally the bias was extreme, with nonsense values returned for factor means and factor variances in some groups. This only occurred in the 15-group conditions. Perhaps large levels of noninvariance interact with number of groups, causing issues of estimation for a few groups in the model. If a researcher is using the alignment primarily to compare factor means, I would recommend ensuring very limited amounts of threshold noninvariance. There were also estimation issues in the complete invariance condition when there were 15 groups, the cause of this is unknown. Again, larger levels of bias were observed primarily in the large magnitude threshold conditions.

Noninvariance Testing Procedure

Overall, the noninvariance testing procedure was too conservative, flagging noninvariant items as invariant a majority of the time, across all conditions. The procedure did not routinely flag noninvariant items in the medium or small conditions and worked better for the thresholds than for the loadings. For the loadings, noninvariant items were found, at best, 65% of the time,

when the magnitude was large, whereas noninvariant thresholds were found 95% of the time for some items. Upon further investigation the method appeared to perform better for items that were the least skewed and had lower MSE. For example, item 10 exhibited the greatest coverage across the latent variable (see Figure 10), and was also found noninvariant the appropriate amount of times. It appears that there is an interaction between the ability of the procedure to estimate the parameter when noninvariance is extreme and the distribution of item responses across the categories, such that if a parameter is poorly estimated due to noninvariance and has skewed thresholds, the procedure cannot detect noninvariance. Further, in the most extreme case, when an item was noninvariant across all groups, the procedure tended to only flag that the parameter was noninvariant in one of the group types.

To better understand where in the procedure issues arise, I further analyzed the creation of the invariance set and the testing of parameters against that invariance set average. From these analyses I concluded that the issue is primarily occurring when the invariance set is created, such that noninvariant parameters are included when they should be excluded. The inclusion of a parameter in the invariance set is determined by a significance test with an alpha value of .001. This appears too conservative, as even large differences were approximately estimated, but not significant at the .001 level. The p-value curves in Figure 11 show that increasing the alpha value will improve the type II error rate substantially. When noninvariant groups are included in the invariance set, the invariant group average is biased, which trickles into the final decision of noninvariance for a given parameter.

Another facet to consider in the noninvariance testing procedure is the suggested effect size measure of noninvariance, R^2. This effect size showed the same pattern of working better for some items than for others. More research needs to be done in the area of the procedure.

From the results of this study, it would appear that any values below .90 should be suspect, as even large magnitudes of noninvariance sometimes had higher R^2 values.

Limitations and Future Research

This study has a number of limitations that could be addressed in future research. First, the use of real data has several important implications. The starting values used in this simulation come from a real large scale noncognitive assessment. Though this adds realism to the simulation, the item specifications were inadvertently manipulated. This made it difficult to tease apart sources of error in the alignment. Future research could include a more simple noninvariance and item specification structure, where item differences are manipulated systematically. These data also restrict the findings to response patterns similar to those seen in noncognitive assessments, where people tend to underutilize the lower categories. Thus, these results may not generalize to items from cognitive tests, which may have a different response pattern. Future research addressing different item types and response patterns is warranted.

In this study I simulated a relatively small number of groups: 3, 9, and 15. The primary reason for this was a lack of computing power. With the resources available to me at the time, conditions with 30 groups took approximately 10 minutes for one replication. I did not find substantial differences in results between 3, 9, and 15 groups, but found some evidence of estimation issues in the 15-group conditions that were not observed in the other conditions. Thus, it is unclear how an increased number of groups influence the results of the alignment. Further, because the group sample size was constant across groups, it is unclear how group sample size and the number of group interact to influence results. This is an important area for future research.

Another area for future research is in further investigating the noninvariance testing procedure. First, I did not manipulate sample size in this study. There is a clear lack of power in the procedure and it would be interesting to see if the hit rate is increased when the sample size is greater than 500. It is unclear, from this study, what the sample size is needed to get the nominal hit rate across all items. A simulation focused on this aspect of the procedure is necessary. Further, it is unclear how the procedure works with continuous or binary items. A comparison is needed to provide further recommendations for the applied researcher. The results here may not generalize to other types of items.

Implications for Practice

As with any new method, best practices need to be established. Though powerful, the automated nature of the procedure creates vulnerability for misuse. In scrolling through the hundreds of pages of output produced by a single alignment run, one is likely to ask themselves, "What do I do with all this information?" Muthén and Asparohouv (2014) suggest conducting a simulation study with real data starting values to evaluate the validity of alignment results. However, for the applied researcher this may not be feasible. Applied researchers need other recommendations for how to interpret the output and evaluate the validity of the results. This is particularly important because as noninvariance increases estimation quality decreases, which causes problems in identifying noninvariant items. There seems to be a bit of a catch-22 with the procedure, in that if more noninvariance occurs, it becomes harder to detect. This issue is not specific to alignment, as the same problem has been noted in other simulation studies with the multiple group method (French & Finch, 2006) The procedure's inability to detect noninvariance seems to be exacerbated when items are skewed in their mean structure, with sparse data in some response categories. For researchers who want to use the alignment method

a simple set of recommendations and criteria needs to be established. As with any statistical method, diagnostics should be considered before results are used to make claims about the data. Given the limitations discussed above I can offer preliminary recommendations that can continue to be evaluated as more research is completed:

1. Researchers need to consider the psychometric properties of their instrument. The alignment appears to work better when item responses are equally distributed across the latent trait, which is less common in noncognitive measurement, as raters tend to utilize the favorable end of the scale. Mplus provides response category frequencies and ICCs, researchers should use this information to examine lack of coverage and sparse data in certain categories.
2. When the alignment completes the model results are listed, then the noninvariance testing. However, researchers should investigate the noninvariance testing output first, as this is the information provided by the program that allows for researchers to evaluate the validity of the alignment model results.
3. Researchers should evaluate the R^2 values for each item, but with caution. Items with a value below .90 are worth further investigation of pairwise differences.
4. They should investigate the estimated pairwise group differences for each parameter and identify differences that are medium or large in magnitude, as they cause bias in factor means and variances. They should use caution if more than 25% of the items have medium or large differences.
5. Because of the conservative nature of the procedure, researchers should not completely rely on the decision printed in the output, but use it in conjunction with the other diagnostic information provided, as is the case with all statistical methods. Generally, if

the R^2 and pairwise differences suggest noninvariance and the item has a skewed response pattern, researchers should interpret the significance test with caution.

6. Researchers may not need to worry about smaller differences, of the magnitude simulated in the small magnitudes conditions. In these conditions there was great coverage and little to no bias.
7. Researchers should use factor variance information with caution, as some estimation issues could occur with larger number of groups, even when there is complete invariance.

There are numerous features that could be added to the software to make following these recommendations easier. First, the invariance testing procedure output would be easier to digest if it were outputted as a dataset. I had to create a Fortran program to do this, which is not easy for an applied researcher. However, with the results in data format, one can plot differences and easily pick out trends. Using the combination of the noninvariance decision, R^2, and estimated pairwise differences researchers have adequate diagnostic information for the model results. Second, the less conservative alpha value for the pairwise tests that create the invariance set should be reconsidered. Perhaps this is only necessary with more complicated models, such as many polytomous items, but from the current simulation it seems too conservative.

This simulation study shows that the alignment is a powerful tool for estimating measurement models across many groups. The method performs well with small and medium levels of noninvariance, allowing for group comparisons across many groups with polytomous items. It appears that the method will have applicability in a large number of situations where researchers have likert type or partial credit items across numerous groups, making comparisons possible that were previously not so.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Joint Committee on Standards for Educational and Psychological Testing.

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 1–14. doi:10.1080/10705511.2014.919210

Baker, E. L. (2013). *Testing in a global future*. Retrieved from http://www.gordoncommission.org/rsc/pdf/baker_testing_global_future.pdf

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated ttem scores: A comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement*, *72*(5), 754–773. doi:10.1177/0013164412440998

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200–223. doi:10.1177/0013164411412768

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 29–51. doi:10.1207/S15327906MBR3601

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 372–398. doi:10.1080/10705511.2012.687671

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures : The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466.

CCSSI. (2015). Standards in your state | Common core state standards initiative. Retrieved January 28, 2015, from http://www.corestandards.org/standards-in-your-state/

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. doi:10.1093/pan/mpt014

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1–27. doi:10.1177/014920639902500101

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of- fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902

Cohen, A. S., Kim, S., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item dunctioning. *Applied Psychological Measurement*, *20*(1), 15–26. doi:10.1177/014662169602000102

Dolan, C. V. (1994). Factor analysis of variable with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*(4), 355–368.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251. doi:10.3102/0013189X15584327

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–32. doi:10.1111/j.1467-8624.2010.01564.x

Ferrer, E., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology (Gott)*, *4*(1), 22–36. doi:10.1027/1614-2241.4.1.22.Factorial

Finch, H. W. (2005). The MIMIC model as a method for detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. doi:10.1177/0146621605275728

Finch, H. W., & French, B. F. (2012). The impact of factor noninvariance on observed score variances. *International Journal of Research and Reviews in Applied Sciences*, *10*(1), 1–13.

Finney, S., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (2nd ed., pp. 439–492). Chartlotte, NC: Information Age Publishing, INC.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology: SOCIAL SCIENCES*, *57B*(5), S275–S284.

French, B. F., & Finch, H. W. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 465–486. doi:10.1207/s15328007sem1303

French, B. F., & Finch, H. W. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 96–113. doi:10.1080/10705510701758349

French, B. F., & Finch, H. W. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, *47*(3), 299–317. doi:10.1111/j.1745-3984.2010.00115.x

French, B. F., & Finch, H. W. (2011). Model Misspecification and Invariance Testing Using Confirmatory Factor Analytic Procedures. *The Journal of Experimental Education*, *79*(4), 404–428. doi:10.1080/00220973.2010.517811

Holland, P. W., & Thayer, D. T. (1986). Diferential item performance and the Mantel-Haenszel procedure. In *Annual Meeting of the American Educational Research Association* (pp. 1–26).

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367. doi:0803973233

Horn, L. J., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*(1), 173–191. doi:10.1007/s11336-003-1136-B

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 642–657. doi:10.1080/10705510903206014

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: Effects of differential item functioning. *The Journals of Gerontology: PSYCHOLOGICAL SCIENCES*, *57*(6), 548–558.

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(409-426).

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639.

Kankaras, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research, 40*(2), 279–310. doi:10.1177/0049124111405301

Kaplan, D. (1989). Power of the likelihood ratio test in multiple group confirmatory factor analysis under partial measurement invariance. *Educational and Psychological Measurement, 49*(3), 579–586.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Lewin, K. (1931). The conflict between Aristotelian and Galileian modes of thought in contemporary psychology. *Journal of General Psychology*, *5*, 141–177.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lubke, G. H., & Muthén, B. O. (2009). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(4), 514–534. doi:10.1207/s15328007sem1104

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. doi:10.1027/1614-1881.1.3.86

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. doi:10.1037//0033-2909.111.3.490

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*. doi:10.1016/0021-9681(79)90031-6

Meade, A. W. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361–388. doi:10.1177/1094428104268027

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635. doi:10.1080/10705510701575461

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*(1), 60–72. doi:10.1207/S15328007SEM1101

Measured Progress, & ETS. (2012). *Smarter Balanced Assessment Consortium General Item Specifications*. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/05/TaskItemSpecifications/ItemSpecifications/GeneralItemSpecifications.pdf

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial ivarinace. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Florence, KY: Routledge.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115. doi:10.1037/1082-989X.9.1.93

Millsap, R. E., & Meredith, W. (2004). Factorial invariance: Historical trends and new developments. In *Factor Analysis at 100*. Chapel Hill, NC: L.L. Thurstone Psychometric Laboratory.

Muthén, B. O. (2002). *Using Mplus monte carlo simulations in practice : A note on assessing estimation quality and power in latent variable models*. *Mplus web notes* (Vol. No. 1).

Muthén, B. O., & Asparouhov, T. (2013a). Item response modeling in Mplus: A multi-dimensional , multi-level, and multi-timepoint example. *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*, 1–29.

Muthén, B. O., & Asparouhov, T. (2013b). *New methods for the study of measurement invariance with many groups*. Retrieved from statmodel.com

Muthén, B. O., & Asparouhov, T. (2014). IRT studies of many groups : The alignment method. *Frontiers in Psychology*, *5*(September), 1–7. doi:10.3389/fpsyg.2014.00978

National Association for the Education of Young Children. (2009). *Where we stand on early learning standards*. Retrieved from https://www.naeyc.org/files/naeyc/file/positions/earlyLearningStandards.pdf

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60. doi:10.1093/pan/mpt014

PARCC. (2015). PARCC task prototypes and new sample items for ELA/literacy. Retrieved January 28, 2015, from http://www.parcconline.org/samples/ELA

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (Second.). Mahwah, NJ: Lawrence Erlbaum Associates.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–66. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8272470

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130*(2), 261–88. doi:10.1037/0033-2909.130.2.261

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. doi:10.1177/0013164413498257

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514.

Schmitt, N., Golubovich, J., & Leong, F. T. L. (2011). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using big five and RIASEC measures. *Assessment*, *18*(4), 412–427. doi:10.1177/1073191110373223

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. doi:10.1016/j.hrmr.2008.03.003

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *The Journal of Applied Psychology*, *91*(6), 1292–1306. doi:10.1037/0021-9010.91.6.1292

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, Inc.*, *25*(1), 78–107.

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, *9*(1), 1–12. doi:10.1027/1614-2241/a000049

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*(1), 118–128. doi:10.1037//0033-2909.99.1.118

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. doi:10.1177/109442810031002

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. doi:10.1177/0146621603259902

Woods, C. M. (2008). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi:10.1177/0146621607314044

Woods, C. M. (2009). Evaluation of MIMIC-Model Methods for DIF Testing With Comparison to Two-Group Analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. doi:10.1080/00273170802620121

Zumbo, B. D., & Koh, K. H. (2005). Manifestation of differences in item-level characteristics in scale-level measurement invariance tests of multi-group confirmatory factor analyses. *Journal of Modern Applied Statistical Methods*, *4*(1), 275–282. Retrieved from http://educ.ubc.ca/faculty/zumbo/Zumbo_Koh_reprint_JMASM.pdf

Appendix A

Example Inputs

3 Groups, 14% NI, Small Magnitude, on the Loadings

```
MONTECARLO:
      NAMES = y1-y14;
      GENERATE= y1-y14 (3);
      !setting scale of variables;
      CATEGORICAL = y1-y14;
      !making categorical;
      NOBSERVATIONS = 3(500);
      NGROUPS = 3;
            NREPS = 500;


   ANALYSIS:
      TYPE =MIXTURE;
      ESTIMATOR = MLR;
      alignment = fixed;
      ALGORITHM=INTEGRATION;
      processors=8;

   MODEL POPULATION:


         %OVERALL%
            TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
              y5*1.006 y6*-.536 y7*.904;

         [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
         [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
         [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
         [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
         [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
         [y6$1*.661 y6$2*1.87 y6$3*2.42];
         [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

           IM by y8*1.715 y9*1.391 y10*.953 y11*.897
               y12*1.253 y13*1.281 y14*1.571;

              [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
         [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
         [y10$1*-.706 y10$2*.85 y10$3*1.892];
         [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
          [y12$1*-3.025 y12$2*-.996 y12$3*.831];
         [y13$1*-2.855 y13$2*-.716 y13$3*.844];
         [y14$1*-3.487 y14$2*-1.252 y14$3*.878];



      %G#1%
```

```
[tm*0];
!factor mean is 0;
    tm*1;
 !factor variance is 1;
[im*0];
 im*1;
tm with im*.56;

            TM by Y1*1.009;
            IM by Y8*1.615;
            !noninvariance for one item on each scale;

%G#2%
[tm*0.3];
    tm*1.5;
[im*0.3];
    im*1.5;
 tm with im*.83;

            TM by Y3*1.187;
            IM by Y10*.853;

            !noninvariance for one item on each scale;

%G#3%
 [tm*1];
    tm*1.2;
 [im*1];
 im*1.2;
 tm with im*.67;

             TM by Y5*.906;
             IM by Y12*1.153;
             !noninvariance for one item on each scale;




model:

 %OVERALL%
    TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
      y5*1.006 y6*-.536 y7*.904;

[y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
[Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
[y3$1*-2.796 y3$2*-1.141 y3$3*.432];
[y4$1*-1.92 y4$2*-.311 y4$3*1.036];
[y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
[y6$1*.661 y6$2*1.87 y6$3*2.42];
```

```
      [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

       IM by y8*1.715 y9*1.391 y10*.953 y11*.897
           y12*1.253 y13*1.281 y14*1.571;

          [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
      [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
      [y10$1*-.706 y10$2*.85 y10$3*1.892];
      [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
       [y12$1*-3.025 y12$2*-.996 y12$3*.831];
      [y13$1*-2.855 y13$2*-.716 y13$3*.844];
      [y14$1*-3.487 y14$2*-1.252 y14$3*.878];
%G#1%
      [tm*0];
      !factor mean is 0;
          tm*1;
       !factor variance is 1;
       [im*0];
       im*1;
      tm with im*.56;

                TM by Y1*1.009;
                IM by Y8*1.615;
                !noninvariance for one item on each scale;

      %G#2%
      [tm*0.3];
          tm*1.5;
       [im*0.3];
          im*1.5;
       tm with im*.83;

                TM by Y3*1.187;
                IM by Y10*.853;

                !noninvariance for one item on each scale;

      %G#3%
       [tm*1];
          tm*1.2;
       [im*1];
       im*1.2;
       tm with im*.67;

                 TM by Y5*.906;
                 IM by Y12*1.153;
                 !noninvariance for one item on each scale;



output: align;
```

3 Groups, 29% NI, Small Magnitude, on the Loadings

```
MONTECARLO:
      NAMES = y1-y14;
      GENERATE= y1-y14 (3);
      !setting scale of variables;
      CATEGORICAL = y1-y14;
      !making categorical;
      NOBSERVATIONS = 3(500);
      NGROUPS = 3;
            NREPS = 500;


   ANALYSIS:
      TYPE =MIXTURE;
      ESTIMATOR = MLR;
      alignment = fixed;
      ALGORITHM=INTEGRATION;
      processors=8;

   MODEL POPULATION:


         %OVERALL%
            TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
              y5*1.006 y6*-.536 y7*.904;

         [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
         [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
         [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
         [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
         [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
         [y6$1*.661 y6$2*1.87 y6$3*2.42];
         [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

           IM by y8*1.715 y9*1.391 y10*.953 y11*.897
               y12*1.253 y13*1.281 y14*1.571;

              [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
         [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
         [y10$1*-.706 y10$2*.85 y10$3*1.892];
         [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
          [y12$1*-3.025 y12$2*-.996 y12$3*.831];
         [y13$1*-2.855 y13$2*-.716 y13$3*.844];
         [y14$1*-3.487 y14$2*-1.252 y14$3*.878];



      %G#1%
        [tm*0];
        !factor mean is 0;
            tm*1;
```

```
  !factor variance is 1;
 [im*0];
 im*1;
tm with im*.56;

        TM by Y1*1.009;
        TM by Y2*.85;
        IM by Y8*1.615;
        IM by Y9*1.291;

        !noninvariance for one item on each scale;

%G#2%
[tm*0.3];
   tm*1.5;
 [im*0.3];
   im*1.5;
 tm with im*.83;

            TM by Y3*1.187;
            TM by Y4*.91;
        IM by Y10*.853;
        IM by Y11*.797;
        !noninvariance for one item on each scale;

%G#3%
 [tm*1];
   tm*1.2;
 [im*1];
 im*1.2;
 tm with im*.67;

        TM by Y5*.906;
        TM by Y6*-.436;
        IM by Y12*1.153;
        IM by Y13*1.181;
        !noninvariance for one item on each scale;



model:
 %OVERALL%
   TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
     y5*1.006 y6*-.536 y7*.904;

[y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
[Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
[y3$1*-2.796 y3$2*-1.141 y3$3*.432];
[y4$1*-1.92 y4$2*-.311 y4$3*1.036];
[y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
```

```
 [y6$1*.661 y6$2*1.87 y6$3*2.42];
 [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

   IM by y8*1.715 y9*1.391 y10*.953 y11*.897
        y12*1.253 y13*1.281 y14*1.571;

      [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
 [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
 [y10$1*-.706 y10$2*.85 y10$3*1.892];
 [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
  [y12$1*-3.025 y12$2*-.996 y12$3*.831];
 [y13$1*-2.855 y13$2*-.716 y13$3*.844];
 [y14$1*-3.487 y14$2*-1.252 y14$3*.878];



%G#1%
 [tm*0];
 !factor mean is 0;
     tm*1;
   !factor variance is 1;
  [im*0];
  im*1;
 tm with im*.56;

              TM by Y1*1.009;
              TM by Y2*.85;
              IM by Y8*1.615;
              IM by Y9*1.291;
              !noninvariance for one item on each scale;

 %G#2%
 [tm*0.3];
     tm*1.5;
  [im*0.3];
     im*1.5;
  tm with im*.83;

                    TM by Y3*1.187;
                    TM by Y4*.91;
              IM by Y10*.853;
              IM by Y11*.797;
              !noninvariance for one item on each scale;

 %G#3%
  [tm*1];
     tm*1.2;
  [im*1];
  im*1.2;
  tm with im*.67;
```

```
        TM by Y5*.906;
        TM by Y6*-.436;
        IM by Y12*1.153;
        IM by Y13*1.181;
        !noninvariance for one item on each scale;


output: align;
```

3 Groups, 14% NI, Small Magnitude, on the Thresholds

```
MONTECARLO:
    NAMES = y1-y14;
    GENERATE= y1-y14 (3);
    !setting scale of variables;
    CATEGORICAL = y1-y14;
    !making categorical;
    NOBSERVATIONS = 3(500);
    NGROUPS = 3;
        NREPS = 500;


 ANALYSIS:
    TYPE =MIXTURE;
    ESTIMATOR = MLR;
    alignment = fixed;
    ALGORITHM=INTEGRATION;
    processors=8;

 MODEL POPULATION:


     %OVERALL%
        TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
          y5*1.006 y6*-.536 y7*.904;

     [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
     [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
     [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
     [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
     [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
     [y6$1*.661 y6$2*1.87 y6$3*2.42];
     [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

      IM by y8*1.715 y9*1.391 y10*.953 y11*.897
          y12*1.253 y13*1.281 y14*1.571;

         [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
     [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
     [y10$1*-.706 y10$2*.85 y10$3*1.892];
     [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
      [y12$1*-3.025 y12$2*-.996 y12$3*.831];
     [y13$1*-2.855 y13$2*-.716 y13$3*.844];
     [y14$1*-3.487 y14$2*-1.252 y14$3*.878];



    %G#1%
     [tm*0];
     !factor mean is 0;
        tm*1;
```

```
  !factor variance is 1;
 [im*0];
 im*1;
tm with im*.56;

            [Y1$1*-3.817 Y1$2*-2.353 Y1$3*-.741];
            [Y8$1*-3.352 Y8$2*-.968 Y8$3*.838];
            !noninvariance for one item on each scale;

%G#2%
[tm*0.3];
    tm*1.5;
 [im*0.3];
    im*1.5;
 tm with im*.83;

            [Y3$1*-2.996 Y3$2*-1.341 Y3$3*.232];
            [Y10$1*-.906 Y10$2*.65 Y10$3*1.692];

            !noninvariance for one item on each scale;

%G#3%
 [tm*1];
    tm*1.2;
 [im*1];
 im*1.2;
 tm with im*.67;
             [Y5$1*-3.484 Y5$2*-2.266 Y5$3*-.797];
             [Y12$1*-3.225 Y12$2*-1.196 Y12$3*.631];

             !noninvariance for one item on each scale;




model:

 %OVERALL%
    TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
      y5*1.006 y6*-.536 y7*.904;

[y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
[Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
[y3$1*-2.796 y3$2*-1.141 y3$3*.432];
[y4$1*-1.92 y4$2*-.311 y4$3*1.036];
[y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
[y6$1*.661 y6$2*1.87 y6$3*2.42];
[y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

 IM by y8*1.715 y9*1.391 y10*.953 y11*.897
```

```
            y12*1.253 y13*1.281 y14*1.571;

          [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
       [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
       [y10$1*-.706 y10$2*.85 y10$3*1.892];
       [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
        [y12$1*-3.025 y12$2*-.996 y12$3*.831];
       [y13$1*-2.855 y13$2*-.716 y13$3*.844];
       [y14$1*-3.487 y14$2*-1.252 y14$3*.878];
  %G#1%
       [tm*0];
       !factor mean is 0;
           tm*1;
         !factor variance is 1;
        [im*0];
        im*1;
       tm with im*.56;

                   [Y1$1*-3.817 Y1$2*-2.353 Y1$3*-.741];
                   [Y8$1*-3.352 Y8$2*-.968 Y8$3*.838];
                   !noninvariance for one item on each scale;

       %G#2%
       [tm*0.3];
           tm*1.5;
        [im*0.3];
           im*1.5;
        tm with im*.83;

                   [Y3$1*-2.996 Y3$2*-1.341 Y3$3*.232];
                   [Y10$1*-.906 Y10$2*.65 Y10$3*1.692];

                   !noninvariance for one item on each scale;

       %G#3%
        [tm*1];
           tm*1.2;
        [im*1];
        im*1.2;
        tm with im*.67;
                    [Y5$1*-3.484 Y5$2*-2.266 Y5$3*-.797];
                    [Y12$1*-3.225 Y12$2*-1.196 Y12$3*.631];

                    !noninvariance for one item on each scale;


output: align;
```

3 Groups, 29% NI, Small Magnitude, on the Thresholds

```
MONTECARLO:
      NAMES = y1-y14;
      GENERATE= y1-y14 (3);
      !setting scale of variables;
      CATEGORICAL = y1-y14;
      !making categorical;
      NOBSERVATIONS = 3(500);
      NGROUPS = 3;
            NREPS = 500;


   ANALYSIS:
      TYPE =MIXTURE;
      ESTIMATOR = MLR;
      alignment = fixed;
      ALGORITHM=INTEGRATION;
      processors=16;

   MODEL POPULATION:


         %OVERALL%
            TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
              y5*1.006 y6*-.536 y7*.904;

         [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
         [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
         [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
         [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
         [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
         [y6$1*.661 y6$2*1.87 y6$3*2.42];
         [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

           IM by y8*1.715 y9*1.391 y10*.953 y11*.897
               y12*1.253 y13*1.281 y14*1.571;

              [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
         [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
         [y10$1*-.706 y10$2*.85 y10$3*1.892];
         [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
          [y12$1*-3.025 y12$2*-.996 y12$3*.831];
         [y13$1*-2.855 y13$2*-.716 y13$3*.844];
         [y14$1*-3.487 y14$2*-1.252 y14$3*.878];



      %G#1%
        [tm*0];
        !factor mean is 0;
            tm*1;
```

```
 !factor variance is 1;
[im*0];
 im*1;
tm with im*.56;

            [Y1$1*-3.817 Y1$2*-2.353 Y1$3*-.741];
            [Y2$1*-1.976 Y2$2*-0.681 Y2$3*0.47];

            [Y8$1*-3.352 Y8$2*-.968 Y8$3*.838];
            [Y9$1*-3.157 Y9$2*-0.976 Y9$3*0.89];

            !noninvariance for two item on each scale;

%G#2%
[tm*0.3];
    tm*1.5;
[im*0.3];
    im*1.5;
 tm with im*.83;

            [Y3$1*-2.996 Y3$2*-1.341 Y3$3*.232];
            [Y4$1*-2.12 Y4$2*-0.511 Y4$3*0.836];

            [Y10$1*-.906 Y10$2*.65 Y10$3*1.692];
            [Y11$1*-2.856 Y11$2*-1.275 Y11$3*0.11];


            !noninvariance for two item on each scale;

%G#3%
 [tm*1];
    tm*1.2;
 [im*1];
 im*1.2;
 tm with im*.67;
            [Y5$1*-3.484 Y5$2*-2.266 Y5$3*-.797];
            [Y6$1*0.461 Y6$2*1.67 Y6$3*2.22];

            [Y12$1*-3.225 Y12$2*-1.196 Y12$3*.631];
            [Y13$1*-3.055 Y13$2*-0.916 Y13$3*0.644];


            !noninvariance for two item on each scale;




model:

 %OVERALL%
```

```
      TM BY y1*1.109 y2*.95 y3*1.287 y4*1.01
        y5*1.006 y6*-.536 y7*.904;

  [y1$1*-3.617 Y1$2*-2.153 Y1$3*-.541];
  [Y2$1*-1.776 Y2$2*-.481 Y2$3*.670];
  [y3$1*-2.796 y3$2*-1.141 y3$3*.432];
  [y4$1*-1.92 y4$2*-.311 y4$3*1.036];
  [y5$1*-3.284 y5$2*-2.066 y5$3*-.597];
  [y6$1*.661 y6$2*1.87 y6$3*2.42];
  [y7$1*-2.939 y7$2*-1.533 y7$3*-.077];

    IM by y8*1.715 y9*1.391 y10*.953 y11*.897
        y12*1.253 y13*1.281 y14*1.571;

      [y8$1*-3.152 y8$2*-.768 y8$3*1.038];
  [y9$1*-2.957 y9$2*-.776 y9$3*1.09];
  [y10$1*-.706 y10$2*.85 y10$3*1.892];
  [y11$1*-2.656 y11$2*-1.075 y11$3*.31];
   [y12$1*-3.025 y12$2*-.996 y12$3*.831];
  [y13$1*-2.855 y13$2*-.716 y13$3*.844];
  [y14$1*-3.487 y14$2*-1.252 y14$3*.878];

%G#1%
  [tm*0];
  !factor mean is 0;
      tm*1;
    !factor variance is 1;
   [im*0];
   im*1;
  tm with im*.56;

              [Y1$1*-3.817 Y1$2*-2.353 Y1$3*-.741];
              [Y2$1*-1.976 Y2$2*-0.681 Y2$3*0.47];

              [Y8$1*-3.352 Y8$2*-.968 Y8$3*.838];
              [Y9$1*-3.157 Y9$2*-0.976 Y9$3*0.89];

              !noninvariance for two item on each scale;

  %G#2%
  [tm*0.3];
      tm*1.5;
   [im*0.3];
      im*1.5;
   tm with im*.83;

              [Y3$1*-2.996 Y3$2*-1.341 Y3$3*.232];
              [Y4$1*-2.12 Y4$2*-0.511 Y4$3*0.836];

              [Y10$1*-.906 Y10$2*.65 Y10$3*1.692];
              [Y11$1*-2.856 Y11$2*-1.275 Y11$3*0.11];
```

```
                !noninvariance for two item on each scale;

        %G#3%
         [tm*1];
            tm*1.2;
         [im*1];
         im*1.2;
         tm with im*.67;
                    [Y5$1*-3.484 Y5$2*-2.266 Y5$3*-.797];
                    [Y6$1*0.461 Y6$2*1.67 Y6$3*2.22];

                    [Y12$1*-3.225 Y12$2*-1.196 Y12$3*.631];
                    [Y13$1*-3.055 Y13$2*-0.916 Y13$3*0.644];


                    !noninvariance for two item on each scale;


output: align;
```

Appendix B

Starting Values for all Conditions

The following tables include the starting values for all conditions. Simulated noninvariance is highlighted in red to show the pattern of noninvariance.

Table B1.

Medium Difference, -.25 from overall except for in the three items condition

		Overall Theta Loadings	One N.I. Item on Each Scale			Two Items			Three Items		
			Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
TM	Y1	1.109	0.859	1.109	1.109	0.859	1.109	1.109	0.859	1.109	1.109
TM	Y2	0.95	0.95	0.95	0.95	0.7	0.95	0.95	0.7	0.95	0.95
TM	Y3	1.287	1.287	1.037	1.287	1.287	1.037	1.287	1.287	1.037	1.287
TM	Y4	1.01	1.01	1.01	1.01	1.01	0.76	1.01	1.01	0.76	1.01
TM	Y5	1.006	1.006	1.006	0.756	1.006	1.006	0.756	1.256	1.006	0.756
TM	Y6	-0.536	-0.536	-0.536	-0.536	-0.536	-0.536	-0.286	-0.536	-0.536	-0.286
TM	Y7	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.654
IM	Y8	1.715	1.465	1.715	1.715	1.465	1.715	1.715	1.465	1.715	1.715
IM	Y9	1.391	1.391	1.391	1.391	1.141	1.391	1.391	1.141	1.391	1.391
IM	Y10	0.953	0.953	0.703	0.953	0.953	0.703	0.953	0.953	0.703	0.953
IM	Y11	0.897	0.897	0.897	0.897	0.897	0.647	0.897	0.897	0.647	0.897
IM	Y12	1.253	1.253	1.253	1.003	1.253	1.253	1.003	1.503	1.253	1.003
IM	Y13	1.281	1.281	1.281	1.281	1.281	1.281	1.031	1.281	1.281	1.031
IM	Y14	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.321

Table B2.

Large Difference, -.40 from overall except for in the three items condition and when there is a negative loading

		Overall Theta Loadings	One N.I. Item on Each Scale			Two Items			Three Items		
			Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
TM	Y1	1.109	0.709	1.109	1.109	0.709	1.109	1.109	0.709	1.109	1.109
TM	Y2	0.95	0.95	0.95	0.95	0.55	0.95	0.95	0.55	0.95	0.95
TM	Y3	1.287	1.287	0.887	1.287	1.287	0.887	1.287	1.287	0.887	1.287
TM	Y4	1.01	1.01	1.01	1.01	1.01	0.61	1.01	1.01	0.61	1.01
TM	Y5	1.006	1.006	1.006	0.606	1.006	1.006	0.606	1.406	1.006	0.606
TM	Y6	-0.536	-0.536	-0.536	-0.536	-0.536	-0.536	-0.136	-0.536	-0.536	-0.136
TM	Y7	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.904	0.504
IM	Y8	1.715	1.315	1.715	1.715	1.315	1.715	1.715	1.315	1.715	1.715
IM	Y9	1.391	1.391	1.391	1.391	0.991	1.391	1.391	0.991	1.391	1.391

IM	Y10	0.953	0.953	0.553	0.953	0.953	0.553	0.953	0.953	0.553	0.953
IM	Y11	0.897	0.897	0.897	0.897	0.897	0.497	0.897	0.897	0.497	0.897
IM	Y12	1.253	1.253	1.253	0.853	1.253	1.253	0.853	1.653	1.253	0.853
IM	Y13	1.281	1.281	1.281	1.281	1.281	1.281	0.881	1.281	1.281	0.881
IM	Y14	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.571	1.171

Table B3.

Medium Difference, -.50 from overall except for in the three items condition

Overall Theta Thresholds			One N.I. Item on Each Scale			Two Items			Three Items		
			Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
TM	Y1$1	-3.617	-4.117	-3.617	-3.617	-4.117	-3.617	-3.617	-4.117	-3.617	-3.617
TM	Y1$2	-2.153	-2.653	-2.153	-2.153	-2.653	-2.153	-2.153	-2.653	-2.153	-2.153
TM	Y1$3	-0.541	-1.041	-0.541	-0.541	-1.041	-0.541	-0.541	-1.041	-0.541	-0.541
TM	Y2$1	-1.776	-1.776	-1.776	-1.776	-2.276	-1.776	-1.776	-2.276	-1.776	-1.776
TM	Y2$2	-0.481	-0.481	-0.481	-0.481	-0.981	-0.481	-0.481	-0.981	-0.481	-0.481
TM	Y2$3	0.67	0.67	0.67	0.67	0.17	0.67	0.67	0.17	0.67	0.67
TM	Y3$1	-2.796	-2.796	-3.296	-2.796	-2.796	-3.296	-2.796	-2.796	-3.296	-2.796
TM	Y3$2	-1.141	-1.141	-1.641	-1.141	-1.141	-1.641	-1.141	-1.141	-1.641	-1.141
TM	Y3$3	0.432	0.432	-0.068	0.432	0.432	-0.068	0.432	0.432	-0.068	0.432
TM	Y4$1	-1.92	-1.92	-1.92	-1.92	-1.92	-2.42	-1.92	-1.92	-2.42	-1.92
TM	Y4$2	-0.311	-0.311	-0.311	-0.311	-0.311	-0.811	-0.311	-0.311	-0.811	-0.311
TM	Y4$3	1.036	1.036	1.036	1.036	1.036	0.536	1.036	1.036	0.536	1.036
TM	Y5$1	-3.284	-3.284	-3.284	-3.784	-3.284	-3.284	-3.784	-2.784	-3.284	-3.784
TM	Y5$2	-2.066	-2.066	-2.066	-2.566	-2.066	-2.066	-2.566	-1.566	-2.066	-2.566
TM	Y5$3	-0.597	-0.597	-0.597	-1.097	-0.597	-0.597	-1.097	-0.097	-0.597	-1.097
TM	Y6$1	0.661	0.661	0.661	0.661	0.661	0.661	0.161	0.661	0.661	0.161
TM	Y6$2	1.87	1.87	1.87	1.87	1.87	1.87	1.37	1.87	1.87	1.37
TM	Y6$3	2.42	2.42	2.42	2.42	2.42	2.42	1.92	2.42	2.42	1.92
TM	Y7$1	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-3.439
TM	Y7$2	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-2.033
TM	Y7$3	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.577
IM	Y8$1	-3.152	-3.652	-3.152	-3.152	-3.652	-3.152	-3.152	-3.652	-3.152	-3.152
IM	Y8$2	-0.768	-1.268	-0.768	-0.768	-1.268	-0.768	-0.768	-1.268	-0.768	-0.768
IM	Y8$3	1.038	0.538	1.038	1.038	0.538	1.038	1.038	0.538	1.038	1.038
IM	Y9$1	-2.957	-2.957	-2.957	-2.957	-3.457	-2.957	-2.957	-3.457	-2.957	-2.957
IM	Y9$2	-0.776	-0.776	-0.776	-0.776	-1.276	-0.776	-0.776	-1.276	-0.776	-0.776
IM	Y9$3	1.09	1.09	1.09	1.09	0.59	1.09	1.09	0.59	1.09	1.09
IM	Y10$1	-0.706	-0.706	-1.206	-0.706	-0.706	-1.206	-0.706	-0.706	-1.206	-0.706
IM	Y10$2	0.85	0.85	0.35	0.85	0.85	0.35	0.85	0.85	0.35	0.85
IM	Y10$3	1.892	1.892	1.392	1.892	1.892	1.392	1.892	1.892	1.392	1.892

IM	Y11$1	-2.656	-2.656	-2.656	-2.656	-2.656	-3.156	-2.656	-2.656	-3.156	-2.656
IM	Y11$2	-1.075	-1.075	-1.075	-1.075	-1.075	-1.575	-1.075	-1.075	-1.575	-1.075
IM	Y11$3	0.31	0.31	0.31	0.31	0.31	-0.19	0.31	0.31	-0.19	0.31
IM	Y12$1	-3.025	-3.025	-3.025	-3.525	-3.025	-3.025	-3.525	-2.525	-3.025	-3.525
IM	Y12$2	-0.996	-0.996	-0.996	-1.496	-0.996	-0.996	-1.496	-0.496	-0.996	-1.496
IM	Y12$3	0.831	0.831	0.831	0.331	0.831	0.831	0.331	1.331	0.831	0.331
IM	Y13$1	-2.855	-2.855	-2.855	-2.855	-2.855	-2.855	-3.355	-2.855	-2.855	-3.355
IM	Y13$2	-0.716	-0.716	-0.716	-0.716	-0.716	-0.716	-1.216	-0.716	-0.716	-1.216
IM	Y13$3	0.844	0.844	0.844	0.844	0.844	0.844	0.344	0.844	0.844	0.344
IM	Y14$1	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.987
IM	Y14$2	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.752
IM	Y14$3	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.378

Table B4.

Large Difference, -.80 from overall except for in the three items condition

	Overall Theta Thresholds		One N.I. Item on Each Scale			Two Items			Three Items		
			Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
TM	Y1$1	-3.617	-4.417	-3.617	-3.617	-4.417	-3.617	-3.617	-4.417	-3.617	-3.617
TM	Y1$2	-2.153	-2.953	-2.153	-2.153	-2.953	-2.153	-2.153	-2.953	-2.153	-2.153
TM	Y1$3	-0.541	-1.341	-0.541	-0.541	-1.341	-0.541	-0.541	-1.341	-0.541	-0.541
TM	Y2$1	-1.776	-1.776	-1.776	-1.776	-2.576	-1.776	-1.776	-2.576	-1.776	-1.776
TM	Y2$2	-0.481	-0.481	-0.481	-0.481	-1.281	-0.481	-0.481	-1.281	-0.481	-0.481
TM	Y2$3	0.67	0.67	0.67	0.67	-0.13	0.67	0.67	-0.13	0.67	0.67
TM	Y3$1	-2.796	-2.796	-3.596	-2.796	-2.796	-3.596	-2.796	-2.796	-3.596	-2.796
TM	Y3$2	-1.141	-1.141	-1.941	-1.141	-1.141	-1.941	-1.141	-1.141	-1.941	-1.141
TM	Y3$3	0.432	0.432	-0.368	0.432	0.432	-0.368	0.432	0.432	-0.368	0.432
TM	Y4$1	-1.92	-1.92	-1.92	-1.92	-1.92	-2.72	-1.92	-1.92	-2.72	-1.92
TM	Y4$2	-0.311	-0.311	-0.311	-0.311	-0.311	-1.111	-0.311	-0.311	-1.111	-0.311
TM	Y4$3	1.036	1.036	1.036	1.036	1.036	0.236	1.036	1.036	0.236	1.036
TM	Y5$1	-3.284	-3.284	-3.284	-4.084	-3.284	-3.284	-4.084	-2.484	-3.284	-4.084
TM	Y5$2	-2.066	-2.066	-2.066	-2.866	-2.066	-2.066	-2.866	-1.266	-2.066	-2.866
TM	Y5$3	-0.597	-0.597	-0.597	-1.397	-0.597	-0.597	-1.397	0.203	-0.597	-1.397
TM	Y6$1	0.661	0.661	0.661	0.661	0.661	0.661	-0.139	0.661	0.661	-0.139
TM	Y6$2	1.87	1.87	1.87	1.87	1.87	1.87	1.07	1.87	1.87	1.07
TM	Y6$3	2.42	2.42	2.42	2.42	2.42	2.42	1.62	2.42	2.42	1.62
TM	Y7$1	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-2.939	-3.739
TM	Y7$2	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-1.533	-2.333
TM	Y7$3	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.077	-0.877
IM	Y8$1	-3.152	-3.952	-3.152	-3.152	-3.952	-3.152	-3.152	-3.952	-3.152	-3.152
IM	Y8$2	-0.768	-1.568	-0.768	-0.768	-1.568	-0.768	-0.768	-1.568	-0.768	-0.768

IM	Y8$3	1.038	0.238	1.038	1.038	0.238	1.038	1.038	0.238	1.038	1.038
IM	Y9$1	-2.957	-2.957	-2.957	-2.957	-3.757	-2.957	-2.957	-3.757	-2.957	-2.957
IM	Y9$2	-0.776	-0.776	-0.776	-0.776	-1.576	-0.776	-0.776	-1.576	-0.776	-0.776
IM	Y9$3	1.09	1.09	1.09	1.09	0.29	1.09	1.09	0.29	1.09	1.09
IM	Y10$1	-0.706	-0.706	-1.506	-0.706	-0.706	-1.506	-0.706	-0.706	-1.506	-0.706
IM	Y10$2	0.85	0.85	0.05	0.85	0.85	0.05	0.85	0.85	0.05	0.85
IM	Y10$3	1.892	1.892	1.092	1.892	1.892	1.092	1.892	1.892	1.092	1.892
IM	Y11$1	-2.656	-2.656	-2.656	-2.656	-2.656	-3.456	-2.656	-2.656	-3.456	-2.656
IM	Y11$2	-1.075	-1.075	-1.075	-1.075	-1.075	-1.875	-1.075	-1.075	-1.875	-1.075
IM	Y11$3	0.31	0.31	0.31	0.31	0.31	-0.49	0.31	0.31	-0.49	0.31
IM	Y12$1	-3.025	-3.025	-3.025	-3.825	-3.025	-3.025	-3.825	-2.225	-3.025	-3.825
IM	Y12$2	-0.996	-0.996	-0.996	-1.796	-0.996	-0.996	-1.796	-0.196	-0.996	-1.796
IM	Y12$3	0.831	0.831	0.831	0.031	0.831	0.831	0.031	1.631	0.831	0.031
IM	Y13$1	-2.855	-2.855	-2.855	-2.855	-2.855	-2.855	-3.655	-2.855	-2.855	-3.655
IM	Y13$2	-0.716	-0.716	-0.716	-0.716	-0.716	-0.716	-1.516	-0.716	-0.716	-1.516
IM	Y13$3	0.844	0.844	0.844	0.844	0.844	0.844	0.044	0.844	0.844	0.044
IM	Y14$1	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-3.487	-4.287
IM	Y14$2	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-1.252	-2.052
IM	Y14$3	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.878	0.078

Appendix C

MSE and Relative Bias of All Thresholds and Loadings for Groups 1-3 in the 26 and 43% Large Noninvariance Conditions

Table C1.
MSE and Relative Bias for all Thresholds in Groups 1, 2, and 3, in the 29% Large, Threshold Noninvariance Conditions, Relative Bias Values Less than -.05 or Greater than .05 are Highlighted

	Average of MSE			Average of Relative Bias			Total Average of MSE	Total Average of Relative Bias
	1	2	3	1	2	3		
Row Average	0.030	0.035	0.166	0.006	-0.014	0.040	0.077	0.011
Y01$1	0.148	0.088	0.503	0.015	0.007	0.067	0.246	0.030
Y01$2	0.054	0.040	0.076	0.010	0.003	0.078	0.057	0.030
Y01$3	0.022	0.027	0.049	0.013	-0.037	0.256	0.033	0.077
Y02$1	0.036	0.031	0.054	0.006	-0.006	0.075	0.040	0.025
Y02$2	0.017	0.022	0.039	0.004	-0.047	0.252	0.026	0.070
Y02$3	0.012	0.024	0.045	0.002	0.045	-0.173	0.027	-0.042
Y03$1	0.042	0.091	0.121	0.008	-0.002	0.070	0.085	0.025
Y03$2	0.018	0.047	0.067	0.008	-0.013	0.148	0.044	0.048
Y03$3	0.014	0.037	0.063	0.008	-0.091	-0.350	0.038	-0.145
Y04$1	0.024	0.046	0.057	0.009	-0.002	0.073	0.042	0.027
Y04$2	0.012	0.026	0.041	0.006	-0.019	0.404	0.026	0.130
Y04$3	0.016	0.026	0.045	0.008	0.111	-0.115	0.029	0.001
Y05$1	0.052	0.059	4.082	0.007	0.001	0.092	1.398	0.033
Y05$2	0.025	0.030	0.111	0.005	-0.008	0.055	0.055	0.017
Y05$3	0.013	0.021	0.058	0.003	-0.038	0.094	0.031	0.020
Y06$1	0.011	0.012	0.022	0.014	-0.013	-0.477	0.015	-0.159
Y06$2	0.022	0.018	0.025	0.012	-0.002	0.068	0.022	0.026
Y06$3	0.031	0.029	0.032	0.013	0.002	0.049	0.030	0.021
Y07$1	0.041	0.047	0.107	0.004	0.003	0.054	0.065	0.021
Y07$2	0.020	0.021	0.041	0.005	-0.005	0.078	0.027	0.026
Y07$3	0.011	0.017	0.034	0.013	-0.258	1.463	0.021	0.406
Y08$1	0.087	0.079	0.117	0.008	0.027	0.033	0.094	0.023
Y08$2	0.027	0.031	0.039	0.007	0.066	0.052	0.032	0.042
Y08$3	0.017	0.035	0.049	0.024	-0.035	-0.013	0.034	-0.008
Y09$1	0.077	0.060	0.093	0.005	0.021	0.028	0.077	0.018
Y09$2	0.022	0.022	0.031	0.008	0.052	0.048	0.025	0.036
Y09$3	0.014	0.027	0.036	-0.024	-0.026	-0.016	0.026	-0.022
Y10$1	0.012	0.021	0.017	0.007	0.014	0.019	0.017	0.014

Y10$2	0.013	0.015	0.019	0.009	-0.467	-0.003	0.016	-0.154
Y10$3	0.021	0.019	0.030	0.009	-0.017	0.002	0.023	-0.002
Y11$1	0.034	0.070	0.058	0.005	0.014	0.020	0.054	0.013
Y11$2	0.014	0.026	0.021	0.005	0.015	0.027	0.020	0.016
Y11$3	0.011	0.014	0.020	-0.018	0.047	-0.050	0.015	-0.007
Y12$1	0.045	0.054	0.220	0.005	0.019	0.035	0.106	0.019
Y12$2	0.016	0.020	0.045	-0.004	0.039	0.018	0.027	0.017
Y12$3	0.016	0.020	0.034	0.007	-0.035	-0.523	0.024	-0.183
Y13$1	0.039	0.051	0.170	0.003	0.019	0.025	0.087	0.016
Y13$2	0.015	0.020	0.042	0.002	0.047	0.016	0.026	0.022
Y13$3	0.016	0.019	0.035	0.009	-0.031	-0.323	0.024	-0.115
Y14$1	0.068	0.069	0.138	0.010	0.018	0.025	0.092	0.018
Y14$2	0.023	0.028	0.034	0.009	0.038	0.029	0.028	0.025
Y14$3	0.018	0.025	0.039	0.006	-0.042	-0.013	0.028	-0.017

Table C2.

MSE and Relative Bias for all Thresholds in Groups 1, 2, and 3, in the 43% Large, Threshold Noninvariance Conditions, Relative Bias Values Less than -.05 or Greater than .05 are Highlighted

	Average of MSE			Average of Relative Bias			Total Average of MSE	Total Average of Relative Bias
	1	2	3	1	2	3		
Row Average	0.029	0.060	0.207	0.006	-0.070	0.273	0.099	0.070
Y01$1	0.147	0.126	0.485	0.015	-0.035	0.051	0.253	0.010
Y01$2	0.053	0.087	0.071	0.009	-0.067	0.051	0.070	-0.002
Y01$3	0.022	0.084	0.047	0.012	-0.318	0.149	0.051	-0.052
Y02$1	0.036	0.069	0.051	0.006	-0.079	0.048	0.052	-0.008
Y02$2	0.017	0.065	0.039	0.004	-0.316	0.152	0.040	-0.054
Y02$3	0.011	0.070	0.045	0.001	0.239	-0.101	0.042	0.046
Y03$1	0.042	0.151	0.103	0.007	-0.051	0.047	0.098	0.001
Y03$2	0.019	0.118	0.056	0.008	-0.105	0.092	0.064	-0.002
Y03$3	0.014	0.118	0.056	0.008	-0.576	-0.204	0.062	-0.257
Y04$1	0.024	0.086	0.049	0.010	-0.053	0.047	0.053	0.001
Y04$2	0.012	0.076	0.033	0.006	-0.146	0.241	0.041	0.034
Y04$3	0.016	0.081	0.038	0.008	0.707	-0.066	0.045	0.216
Y05$1	0.031	0.108	4.069	0.005	-0.040	0.079	1.403	0.014
Y05$2	0.016	0.084	0.105	0.003	-0.074	0.036	0.068	-0.012
Y05$3	0.012	0.074	0.056	0.004	-0.266	0.056	0.047	-0.069
Y06$1	0.011	0.024	0.022	0.014	-0.123	-0.281	0.019	-0.130
Y06$2	0.022	0.028	0.025	0.012	-0.040	0.043	0.025	0.005

Y06$3	0.031	0.035	0.031	0.013	-0.028	0.032	0.032	0.006
Y07$1	0.041	0.077	0.233	0.004	-0.038	0.041	0.117	0.002
Y07$2	0.020	0.057	0.072	0.005	-0.085	0.038	0.050	-0.014
Y07$3	0.011	0.057	0.046	0.013	-1.851	0.074	0.038	-0.588
Y08$1	0.088	0.078	0.146	0.008	0.003	-0.040	0.104	-0.010
Y08$2	0.027	0.039	0.098	0.007	-0.032	-0.248	0.054	-0.091
Y08$3	0.017	0.046	0.125	0.025	0.038	0.210	0.063	0.091
Y09$1	0.076	0.063	0.105	0.005	0.000	-0.034	0.081	-0.010
Y09$2	0.022	0.028	0.067	0.007	-0.026	-0.192	0.039	-0.070
Y09$3	0.014	0.034	0.082	-0.024	0.030	0.154	0.044	0.053
Y10$1	0.012	0.024	0.037	0.007	-0.014	-0.161	0.025	-0.056
Y10$2	0.013	0.019	0.042	0.009	0.385	0.147	0.025	0.180
Y10$3	0.021	0.025	0.055	0.009	0.022	0.070	0.034	0.034
Y11$1	0.034	0.069	0.059	0.005	0.003	-0.024	0.054	-0.005
Y11$2	0.014	0.028	0.031	0.006	-0.006	-0.082	0.024	-0.027
Y11$3	0.011	0.017	0.035	-0.018	-0.035	0.330	0.021	0.093
Y12$1	0.029	0.060	0.220	0.003	0.000	-0.010	0.103	-0.002
Y12$2	0.014	0.028	0.078	-0.021	-0.017	-0.076	0.040	-0.038
Y12$3	0.021	0.029	0.076	0.001	0.031	4.941	0.042	1.658
Y13$1	0.039	0.048	0.184	0.003	0.000	-0.023	0.090	-0.007
Y13$2	0.015	0.023	0.082	0.002	-0.031	-0.099	0.040	-0.043
Y13$3	0.016	0.023	0.083	0.008	0.035	3.656	0.041	1.233
Y14$1	0.067	0.074	1.156	0.010	-0.002	-0.010	0.432	-0.001
Y14$2	0.023	0.035	0.111	0.009	-0.017	-0.083	0.056	-0.030
Y14$3	0.018	0.034	0.107	0.006	0.036	2.426	0.053	0.822

Table C3.

MSE and Relative Bias for all Loadings in Groups 1, 2, and 3, in the 29 and 43% Large, Loading Noninvariance Conditions, Relative Bias Values Less than -.05 or Greater than .05 are Highlighted

	Average of MSE			Average of Relative Bias			Total Average of MSE	Total Average of Relative Bias
Row Average	1	2	3	1	2	3		
29%	0.022	0.024	0.029	0.008	0.009	0.001	0.025	0.006
Y01	0.022	0.038	0.044	0.008	0.032	-0.008	0.035	0.011
Y02	0.014	0.028	0.030	0.005	0.026	-0.005	0.024	0.008
Y03	0.033	0.026	0.050	0.012	0.028	0.000	0.036	0.013
Y04	0.023	0.017	0.029	0.007	0.027	-0.004	0.023	0.010
Y05	0.026	0.032	0.024	0.006	0.017	0.008	0.027	0.010
Y06	0.017	0.012	0.012	0.013	0.008	-0.034	0.014	-0.004

Y07	0.021	0.023	0.022	0.010	0.027	-0.007	0.022	0.010
Y08	0.023	0.042	0.058	0.009	-0.003	0.015	0.041	0.007
Y09	0.016	0.026	0.035	0.001	-0.004	0.009	0.026	0.002
Y10	0.016	0.009	0.016	0.012	-0.009	0.014	0.014	0.006
Y11	0.016	0.009	0.016	0.003	-0.003	-0.006	0.014	-0.002
Y12	0.020	0.020	0.019	0.005	-0.004	0.011	0.020	0.004
Y13	0.023	0.020	0.021	0.012	-0.007	0.008	0.021	0.004
Y14	0.032	0.027	0.034	0.011	-0.004	0.011	0.031	0.006
43%	0.023	0.036	0.052	0.008	0.075	0.106	0.037	0.063
Y01	0.021	0.062	0.076	0.007	0.114	0.103	0.053	0.075
Y02	0.013	0.045	0.052	0.005	0.107	0.104	0.037	0.072
Y03	0.030	0.040	0.090	0.010	0.109	0.110	0.053	0.077
Y04	0.021	0.025	0.053	0.007	0.109	0.104	0.033	0.073
Y05	0.042	0.051	0.039	0.008	0.100	0.120	0.044	0.076
Y06	0.016	0.016	0.016	0.014	0.087	0.084	0.016	0.062
Y07	0.020	0.040	0.025	0.008	0.109	0.109	0.028	0.075
Y08	0.022	0.057	0.119	0.009	0.047	0.114	0.066	0.057
Y09	0.015	0.037	0.070	0.001	0.046	0.105	0.041	0.051
Y10	0.015	0.012	0.034	0.012	0.042	0.112	0.020	0.055
Y11	0.015	0.011	0.027	0.004	0.048	0.091	0.018	0.047
Y12	0.032	0.033	0.035	0.006	0.048	0.109	0.033	0.054
Y13	0.021	0.027	0.035	0.011	0.043	0.105	0.028	0.053
Y14	0.030	0.041	0.053	0.009	0.046	0.109	0.041	0.055

Appendix D

Hit Rate for All Thresholds in the 29-43% Large Magnitude Threshold Noninvariance Conditions

Table D1.
Percentage of Times across Replications that a Parameter was Flagged as Noninvariant in the 29 and 43% Large Threshold Noninvariance Conditions, Noninvariant Parameters are Highlighted

	29% Noninvariance			43% Noninvariance		
	Group Type					
Parameter	1	2	3	1	2	3
Item 1	**29.2**	**6.2**	**0.9**	**33.0**	**7.7**	**0.9**
Threshold 1	0.3	4.2	1.0	0.8	6.6	1.0
Threshold 2	9.6	10.0	1.0	16.8	12.8	1.2
Threshold 3	77.7	4.4	0.6	81.5	3.7	0.5
Item 2	**75.3**	**3.7**	**0.6**	**78.8**	**4.2**	**0.9**
Threshold 1	39.6	9.6	1.5	52.1	9.8	1.6
Threshold 2	91.7	1.0	0.1	91.0	1.8	0.5
Threshold 3	94.6	0.6	0.2	93.2	1.1	0.6
Item 3	**5.7**	**34.9**	**0.5**	**1.5**	**13.5**	**0.2**
Threshold 1	5.4	1.8	0.7	1.4	0.9	0.4
Threshold 2	9.6	31.0	0.6	2.3	11.9	0.2
Threshold 3	2.0	72.0	0.2	0.9	27.8	0.0
Item 4	**4.4**	**63.2**	**0.6**	**2.0**	**32.1**	**0.4**
Threshold 1	9.4	19.6	1.4	3.7	7.7	0.9
Threshold 2	2.9	84.6	0.4	1.6	43.9	0.1
Threshold 3	0.8	85.4	0.0	0.7	44.6	0.0
Item 5	**8.5**	**9.4**	**36.6**	**66.2**	**35.1**	**32.5**
Threshold 1	4.7	4.9	0.5	84.7	17.3	2.3
Threshold 2	19.5	21.7	16.9	78.8	46.9	20.9
Threshold 3	1.3	1.5	92.5	35.1	41.2	74.2
Item 6	**0.4**	**0.3**	**86.3**	**1.9**	**0.5**	**79.1**
Threshold 1	0.6	0.6	95.5	4.0	1.5	87.7
Threshold 2	0.2	0.1	89.9	0.8	0.0	82.8
Threshold 3	0.2	0.0	73.6	0.7	0.1	66.7
Item 7	**0.1**	**0.1**	**0.1**	6.1	9.9	47.0
Threshold 1	0.1	0.2	0.0	6.2	12.1	0.6
Threshold 2	0.1	0.2	0.0	11.4	15.7	51.9
Threshold 3	0.0	0.0	0.3	0.8	1.8	88.5
Item 8	**56.9**	**1.2**	**0.7**	**61.8**	**0.6**	**1.0**
Threshold 1	2.5	2.1	1.6	7.5	1.5	2.7
Threshold 2	76.4	1.4	0.4	86.0	0.2	0.2

Threshold 3	91.7	0.1	0.1	91.9	0.1	0.1
Item 9	**64.7**	**1.6**	**0.5**	**67.1**	**1.1**	**0.9**
Threshold 1	5.7	4.3	1.4	12.2	3.0	2.5
Threshold 2	91.5	0.3	0.1	93.3	0.1	0.2
Threshold 3	97.0	0.2	0.1	95.9	0.2	0.1
Item10	**0.2**	**98.8**	**0.1**	**0.1**	**97.4**	**0.1**
Threshold 1	0.4	98.7	0.2	0.1	97.6	0.2
Threshold 2	0.1	99.5	0.1	0.0	98.4	0.0
Threshold 3	0.1	98.1	0.0	0.1	96.2	0.0
Item11	**3.0**	**68.9**	**1.2**	**1.3**	**69.0**	**1.4**
Threshold 1	7.6	10.9	3.1	3.7	12.2	4.0
Threshold 2	1.4	95.9	0.4	0.1	96.3	0.1
Threshold 3	0.0	99.8	0.1	0.1	98.6	0.1
Item12	**6.6**	**5.3**	**47.8**	**87.9**	**30.1**	**13.0**
Threshold 1	5.6	4.5	0.2	91.1	8.8	0.7
Threshold 2	13.4	11.2	59.6	90.1	45.2	13.7
Threshold 3	0.7	0.4	83.6	82.4	36.4	24.7
Item13	**4.7**	**3.6**	**52.5**	**2.5**	**3.9**	**22.0**
Threshold 1	7.2	5.0	0.2	2.6	3.2	0.0
Threshold 2	6.2	5.5	74.1	4.5	7.3	27.8
Threshold 3	0.7	0.4	83.1	0.4	1.3	38.0
Item14	**0.1**	**0.0**	**0.1**	**1.5**	**2.4**	**11.6**
Threshold 1	0.2	0.0	0.1	1.1	1.4	0.4
Threshold 2	0.0	0.0	0.0	3.0	5.3	9.6
Threshold 3	0.0	0.0	0.1	0.3	0.6	24.7